META-ANALYSIS OF THE DIAGNOSTIC ACCURACY OF THE SCHOOL-AGE ACHENBACH SYSTEM OF EMPIRICALLY-BASED ASSESSMENT (ASEBA)

Jacquelynne Genzlinger

A thesis submitted to the faculty at the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Master of Arts in the Department of Psychology and Neuroscience in the College of Arts and Sciences (Clinical Psychology).

Chapel Hill
2019

Approved by:

Mitch Prinstein

Jennifer K. Youngstrom

Eric A. Youngstrom

ABSTRACT

Jacquelynne Genzlinger: Meta-Analysis of the Diagnostic Accuracy of the School-Age Achenbach System of Empirically Based Assessment (ASEBA)
(Under the direction of Eric Youngstrom)

Achenbach System of Empirically-Based Assessment (ASEBA) is a popular set of instruments used to detect problem behaviors in youths. Diagnostic accuracy is the ability of a test to discriminate between those who have a target condition and those who do not. The present study is a comprehensive review and multivariate meta-analysis of studies evaluating the diagnostic accuracy of ASEBA for any psychiatric disorder in children 6-18 using sensitivity and specificity or area under the curve (AUC) analysis. Moderating variables include informant, diagnosis, and study design characteristics. The analysis included 223 unique effect sizes from 13,516 youths across 28 studies. The average pooled effect size was large, g = 1.02, indicating ASEBA showed good discriminative validity overall. Caregiver report performed significantly better than teacher or youth report. ASEBA showed better discrimination for PBD, CD, and ODD compared to ADHD. Findings support continued use of ASEBA for discriminating psychological disorders in youth.

TABLE OF CONTENTS

LIST OF FIGURES vi

LIST OF TABLES vii

LIST OF ABBREVIATIONS viii

Introduction 1

Advantages of the ASEBA Approach 1

Criticisms of ASEBA 3

Diagnostic Accuracy 5

Specific Aims & Research Questions 7

Methods 9

Search Protocol 9

Inclusion and Exclusion Criteria 9

Coding Protocol 10

Operational Definitions of Coding Constructs 11

Measures Included 13

Data Analytic Plan 14

Results 16

Assessment of Study Quality 17

Overall Summary of Effect Sizes 17

Simple Moderator Analyses 18

Main Model: Multivariate Meta-Regression with All Predictors 20

Testing the Robustness of the Meta-Regression Models 20

Clinical Interpretability 21

Discussion 22

Moderators of Diagnostic Accuracy Effect Sizes 22

Consideration of Alternative Explanations 26

Generalizability of Conclusions 27

Limitations 27

Future Directions 28

REFERENCES 53

LIST OF FIGURES

Figure 1. Diagram of ASEBA Syndrome Scales 29

Figure 2. Flow diagram of included studies.. 30

Figure 3. Forest plot for ADHD. 31

Figure 4. Forest plot for anxiety disorders. 32

Figure 5. Forest plot for bipolar spectrum disorders... 33

Figure 6. Forest plot for conduct disorder. 34

Figure 7. Forest plot for depressive disorders. 35

Figure 8. Forest plot for externalizing disorders. 36

Figure 9. Forest plot for internalizing disorders. 37

Figure 10. Forest plot for oppositional defiant disorder. 38

Figure 11. Forest plot for post-traumatic stress disorder. 39

Figure 12. Forest plot for substance misuse disorders. 40

LIST OF TABLES

Table 1. Sample-Level Characteristics 41

Table 2. Effect Size Level Characteristics and Moderators. 49

Table 3. Simple Moderator Analysis 50

Table 4. Fully Augmented Model 51

Table 5. Predicted Effect Sizes 52

LIST OF ABBREVIATIONS

ADHD	Attention-Deficit/Hyperactivity Disorder
ASEBA	Achenbach System of Empirically Based Assessment
AUC	Area Under the Curve
CBCL	Child Behavior Checklist
CD	Conduct Disorder
DSM	Diagnostic and Statistical Manual of Mental Disorders
ODD	Oppositional Defiant Disorder
PBD	Pediatric Bipolar Disorder
PTSD	Post-traumatic Stress Disorder
ROC	Receiver Operating Characteristic Curve
TRF	Teacher Report Form
YSR	Youth Self Report

Introduction

The school-age Achenbach System of Empirically-Based Assessment (ASEBA) is a commonly-used, low-cost collection of instruments used for the detection of problem behaviors in children and teens aged 6-18 (Achenbach, 1991; Achenbach & Edelbrock, 1983; Achenbach & Rescorla, 2001). ASEBA includes a caregiver report form called the Child Behavior Checklist (CBCL), a checklist to be completed by the child’s teacher(s) (Teacher Report Form, or TRF), as well as a self-report form for youths aged 11-18 (Youth Self Report, or YSR). Factor analyses have broken the items into eight domains of problem areas, called the empirically-based syndrome scales: Aggressive Behavior, Anxious/Depressed, Attention Problems, Rule-Breaking Behavior, Somatic Complaints, Social Problems, Thought Problems, and Withdrawn/Depressed (Achenbach & Rescorla, 2001; Lengua et al., 2001). Figure 1 shows the how these syndrome scales combine to form a total score, an internalizing problems score, and an externalizing problems score. More recently, a panel of experts have identified ASEBA items that are consistent with diagnostic criteria in the DSM-5 in order to create DSM-oriented scales (Achenbach & Dumenci, 2001). These six scales include: Depressive Problems, Anxiety Problems, Somatic Problems, Attention Deficit/Hyperactivity Problems, Oppositional Defiant Problems, and Conduct Problems.

Advantages of the ASEBA Approach

Ease of Administration. The checklists are at a sixth-grade reading level, which allows ASEBA to be used with a wide range of individuals regardless of educational status or verbal ability. The estimated completion time for each form is about 13 minutes for the CBCL and 8-10 minutes for the TRF (Camara, Nathan, & Puente, 2000). This makes it convenient for caregiver(s), teacher(s), and/or the child to complete the questionnaires quickly. ASEBA is also a low-cost, affordable option. At present, ASEBA costs sixty cents per form -- although it requires the additional purchase of the interpretation manual and scoring materials (which include the choice of hand-scoring forms and templates, computer scoring software, or a subscription to score online).

Multi-informant assessment. The context that children are in has a large impact on how they behave. Achenbach is a pioneer for creating one of the first collections of child assessments that utilize information from multiple informants to account for situational specificity, or the idea that different people have different perspectives of the child's behavior depending on the context (Achenbach, 1995; De Lose Reyes & Kazdin, 2005). A teacher will be able to give descriptions of how the child behaves in the classroom, and this may be different from how this child acts for caregivers within the home. Gathering information from multiple people will give a more complete picture of the child's behavior in many domains; this is important for a clinician to know when considering a child's diagnosis and treatment.

Norms. Norms are created by choosing samples that represent an overall population; the distribution of scores within the standardization sample is thought to be a good estimate of how the scores would be spread within the larger population of interest. The most recent norms of ASEBA were from a non-referred sample of 1,753 children across 40 states (Achenbach & Rescorla, 2001). A child's score is compared to the observed scores of the children in the standardization sample who are the same age and sex of that child in order to determine if the child's score is normal, borderline-clinical, or clinically significant. Knowing a child's score in comparison to their peers can be very helpful in understanding the appropriateness of their behaviors and the severity of the child's issues, and is important for practitioners to consider when deciding whether the child needs treatment (Achenbach, 2001).

Criticisms of ASEBA

Multicultural Issues. Although ASEBA has been translated into over 100 languages, critics have raised concerns over the multicultural sensitivity of these forms. A literal translation of words in a measure from one language into another does not ensure that the symptoms will be interpreted in the same way by respondents of other cultural and ethnic groups (Kleinman & Good, 1985). Parents sometimes interpret their children's symptoms in different ways (Harkness & Super, 1990) or might may have different threshold for distress or concerns about particular behavioral problems depending on their own cultural background. ASEBA was standardized using in the U.S. population, so these norms may not be equally as applicable to diverse cultures, ethnicities, and socioeconomic groups that were poorly represented in the standardization sample (Bird, Gould, Rubio-Stipec, Staghezza, & Canino, 1991). The base rates of behavioral symptoms can vary within children from different groups (Gopalkrishnan & Babacan, 2015), and the perceived cause of mental illness has been shown to differ across cultures (Choudhry, Mani, Ming, & Khan, 2016).

Missing Items for Certain Childhood Disorders. ASEBA was created in a time when bipolar disorder was still conceptualized as an adult phenomenon and not accepted as occurring in pediatric populations yet. As a result, ASEBA is missing questions regarding mania-specific symptoms, such as grandiosity or elated mood, and the manic symptoms that are included could also be attributed to other conditions (Youngstrom, Genzlinger, Egerton, & Van Meter, 2015). Some have tried to create a "bipolar profile" by combining items from different scales ASEBA; however, further research did not provide support for its use and found that it did not have much incremental value beyond the Externalizing score (Diler et al., 2009; Meyer et al., 2009; Kahana, Youngstrom, Findling, & Calabrese, 2003, 2004; Youngstrom et al., 2015). Additionally, ASEBA has been found to be inadequate in assessing for autism for similar reasons (Havdahl et al., 2016).

Use of Norms Instead of Comparison Groups. The normative sample consisted of 44% males, 33% upper class, 51% middle class, 60% Caucasian, 20% African American, and 9% Latino children (Achenbach & Rescorla, 2001). Children were excluded from the normative sample for a variety of reasons, including presence of an intellectual disability, serious illness, or disability, as well as if their parents did not speak English, or if the children had any contact with a mental health or substance abuse service within the past year (Achenbach & Rescorla, 2001). This may not be the most helpful or ideal comparison depending on if the target sample or case has different characteristics than the individuals ASEBA was normed against.

Combining Data Across Informants. The inclusion of collateral information from multiple informants is a strength of ASEBA. However, when different informants endorse varying levels of symptoms or even opposing views on an issue, how to interpret this information can become tricky. One might be tempted to favor self-report as more accurate and disregard the caregiver information, or find the parent more trustworthy and ignore what the teen has to say. Neither of these decisions would be optimal. A meta-analysis by Achenbach and colleges (1987) argues that we cannot replace the information from one informant with data from another informant, as observations about a single child can vary greatly based on the situational context.

So, if we should not disregard a piece of information or try to replace it with data from another informant, what should we do? There are Bayesian methods from an evidence-based assessment approach that allow us to actually combine the information of multiple informants statistically through the use of a probability nomogram (Youngstrom, Choukas-Bradley, Calhoun, & Jensen Doss, 2014). A person starts out with a base rate probability of a disorder, and the test score can be converted into a likelihood ratio that can be plugged in to the nomogram to yield an updated probability based on the combination of the starting probability and the test score. It is possible to input multiple test scores from different informants so that the probability can reflect all these pieces of information. The clinician will not have to decide which informant to weigh more heavily or which piece of information is better, and it is more unbiased to use this kind of actuarial method.

Diagnostic Accuracy

Diagnostic accuracy is the ability of a test to discriminate between those who have a target condition and those who do not. How well a test performs is not a fixed attribute of the measure itself, but depends on other factors such as the prevalence of the target condition, how one defines the disorder and whom to include in the target group according to that definition (e.g., spectrum or subthreshold diagnoses, comorbidities), and how the study testing the measure was designed. In a perfect world, a test would always be able to distinguish between those with and without a disorder. However, this is not what we actually see happen in practice. There are four possible outcomes of how a person is classified by the test (see table below). The measure can correctly identify a person who has the condition as having the condition (true positive), or it can correctly determine that a person without the condition does not have the condition (true negative). The test can also make errors, where it incorrectly classifies an individual with the condition as not having the condition (false negative), or incorrectly categorize the person who does not have the condition as having the condition (false positive).

	Person has target condition	**Person does not have target condition**
Positive test	True Positive	False Positive
Negative test	False Negative	True Negative

The *sensitivity* of a measure is the probability of getting a positive test result in subjects who actually have the disorder. In other words, if 100 people had the target condition, how many of those people were positively identified by the test as having the condition? The formula for sensitivity is:

$$\text{Sensitivity} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}}$$

On the flip side, *specificity* is the probability that individuals who do not have the condition will get a negative test result. If 100 people without the condition were studied, how many were correctly identified as not having the disorder by receiving a negative result on the test? The formula for this is:

$$\text{Specificity} = \frac{\text{True Negative}}{\text{True Negative} + \text{False Positive}}$$

Sensitivity and specificity pairs at each cutoff point of the test can be plotted to create a receiver operating characteristic curve (ROC) graph. The area under the ROC curve (AUC) is an overall measure of the diagnostic accuracy of the test averaged across all the sensitivity-specificity pairs. An AUC can range from 0-1, where a value of 0.5 would indicate the test is performing at chance, and an AUC of 1 would indicate perfect discrimination. The larger the AUC is, the better the test is at differentiating between those who have the disorder and those who do not.

Specific Aims & Research Questions

The current study is a systematic review and meta-analysis of the diagnostic efficiency of ASEBA for any psychiatric disorder in youths aged 6-18. The main research questions are:

(a) Is ASEBA an effective method of assessing for the presence of common psychiatric disorders in children?

The present study will consolidate the sensitivity-specificity pairs or AUC values that are reported in published research in order to evaluate whether ASEBA is an effective method to assess for the presence of common psychological conditions in children. The higher the AUC values are across all of the included studies, the larger the effect size will be, and the more effective we can conclude that ASEBA is for this purpose.

(b) Does the type of informant change the diagnostic accuracy of ASEBA?

Although data from multiple informants can be combined through Bayesian methods (discussed earlier), there may be specific instances where it could be less helpful to give equal weight to all informants. For example, relying on a youth's self-report may obscure the assessment of externalizing behaviors because the adolescent may lack insight or not be willing to acknowledge there is a problem. Additionally, self-report may be more valuable for internalizing disorders specifically, because the teen will have better awareness of what he/she is feeling internally that the caregiver may not know about (Loeber, Green, & Lahey, 1990). When compared head to head, I expect that the CBCL will have the largest effect sizes of diagnostic accuracy over the YSR and TRF, as has been found in a prior meta-analysis that focused more narrowly on bipolar disorder (Youngstrom et al., 2015). I hypothesize that the YSR will have the largest effect sizes for anxiety and depression if we compare how it performs across multiple diagnoses. I believe the TRF will have poor diagnostic accuracy overall, with the exception of displaying relatively larger effect sizes for externalizing behaviors such as ADHD, ODD, and CD.

(c) Is ASEBA better suited for detecting certain types of childhood disorders over others?

The present study is not narrowly focused on a single disorder, nor is it purposely excluding any diagnoses – any study that fit the inclusion criteria was analyzed. Due to higher prevalence of certain disorders in childhood, we expect to find studies that evaluate ASEBA in depression, anxiety, attention-deficit hyperactivity disorder (ADHD), oppositional defiant disorder (ODD), conduct disorder (CD), and pediatric bipolar disorder (PBD) (Merikangas et al., 2011). Prior research has shown that the CBCL has limited ability to predict depression and is less able to differentiate between types of internalizing problems (Song et al., 1994; Mash & Barkley, 2014), and as discussed previously, is suboptimal for the assessment of pediatric bipolar disorder. I expect that ASEBA will have larger effect sizes for disruptive behavior disorders such as ADHD, ODD, and CD, and smaller effect sizes for anxiety, depression, and bipolar disorder.

Methods

Search Protocol

A reference librarian who specializes in systematic searches was consulted in order to optimize the searching strategies. The comprehensive literature search was conducted in late fall of 2018 using PsycINFO, PubMed, and TRIP. The search terminology were: (("child behavior checklist" OR CBCL) OR (“youth self report” OR YSR) OR (“teacher report form” OR TRF) OR (“Achenbach system of empirically based assessment” OR ASEBA)) AND ((sensitivity AND specificity) OR “diagnostic efficiency” or “diagnostic accuracy” OR “area under the curve” OR AUC OR “receiver operating characteristic” OR ROC). Searches in October 2018 yielded a total of 373 hits in PsycINFO, 122 hits in PubMed, and 105 hits in TRIP. The titles and abstracts of all hits from each database were reviewed and articles that appeared to provide the appropriate relevant information were downloaded and further evaluated using the inclusion and exclusion criteria for this study. Figure 1 is a flow diagram of the search process.

Inclusion and Exclusion Criteria

Articles included in the study needed to: evaluate the diagnostic efficiency of ASEBA for a psychiatric disorder in children aged 6-18 years against a valid reference standard (i.e., a structured diagnostic interview) to identify cases with the target disorder. These articles also needed to include sufficient data for an effect size of diagnostic efficiency to be calculated. The options of statistics that could be reported include both the sensitivity and specificity of the measure or an area under the curve (AUC) or receiver operating characteristic (ROC) curve analysis of its performance. Articles were excluded if: not published in English, did not include sufficient data to calculate an effect size, or included children younger than age 5 or adults older than age 18. Diagnoses that were derived without a valid reference standard (e.g., clinical diagnoses using an unstructured interview, obtaining diagnoses through chart review) were also be excluded, in addition to any data that reports a pooled estimate for all psychiatric diagnoses combined instead of separating estimates per each disorder.

Coding Protocol

A group of four senior undergraduate and post-baccalaureate psychology majors were supervised and trained in the coding protocol by a doctoral student. All coders initially met as a group to learn how to convert the reported statistics into the desired effect sizes and how to code information from articles using the coding manual. In order to be able to code independently, each person needed to code six articles on her own after the initial training meeting, then met once more as a group to ensure that these articles were coded correctly and discuss any new questions with the graduate student mentor.

The coding manual was an Excel file where coders filled in cells with relevant information from each article, including: year of publication, year(s) of data collection, country of data collection, whether the measure was translated (and the language it was translated to, if yes), the setting where the data were collected, and who funded the study. Demographic information about participants was also coded, including age, percent female, percent white (as an estimate of how diverse the sample is), socioeconomic status, and percentage of participants with various psychiatric disorders (e.g., ADHD, depression, anxiety) as well as percent healthy control. Also coded were specific characteristics of the study, including time lag between when the ASEBA measure was completed and when the diagnostic interview was conducted (which is relevant for diagnoses that are episodic), the type reference standard used, attrition, and if the sampling was distilled. Coders recorded the reported sensitivity and specificity at a specific threshold or cut score, and/or the AUC of the measure from the study. Finally, quality of the study design and reporting was coded using criteria from QUADAS-2 (Whiting et al., 2011) and Kowatch (Kowatch et al., 2005), consistent with other recent papers evaluating diagnostic efficiency (Youngstrom et al., 2015; Youngstrom et al., 2018).

The first set of coding took place during the fall of 2015 for articles identified by the preliminary searches. The updated search during fall of 2018 identified additional articles published since 2015 that met inclusions criteria for the present study. Each article was independently coded by two individuals for study characteristics, effect size, moderator variables, and quality variables. The graduate student supervisor consolidated the double coding and independently resolved discrepancies between any conflicting codes. When any discrepancy remained unsolved by the graduate student, the faculty mentor was consulted for final resolution.

Operational Definitions of Coding Constructs

Sample Characteristics. *Distilled sample* – This was a dummy-coded, dichotomous variable to indicate whether the study design was likely to inflate effect sizes through the inclusion of healthy controls or the use of narrow inclusion criteria that would eliminate other diagnoses or comorbidities similar to the target condition that would make it harder for the measure to distinguish from one another (Bossuyt et al., 2003; Youngstrom et al., 2006; Youngstrom et al., 2015).

Target diagnosis in study – The present analysis included studies that examine the performance of ASEBA across a variety of diagnoses. The target diagnosis variable was used to quantify which disorder a study was discriminating from the rest of the sample through the use of ASEBA.

Definition of diagnosis in study – This variable was dummy-coded to compare broad versus narrow definitions of the target diagnosis. Using depression as an example, major depressive disorder would be considered a narrow definition because it is only one of the depressive disorders. The use of multiple depressive disorders (such as dysthymic disorder and minor depression in addition to major depressive disorder) would qualify as a broad definition of the target diagnosis. Studies using a broad definition will capture a range of severity by including spectrum disorders, which will likely result in lower effect size estimates on average compared to those using narrow definitions.

Clinical Setting. The setting where data collection took place was coded based on the following levels of severity: (a) inpatient, hospitalized, or residential, (b) specialty outpatient clinic or research study, (c) outpatient, (d) at risk, or (e) general community. Whenever a paper included data from more than one type of setting, coders assigned the study the code associated with the more severe setting. More acute settings will likely yield higher overall scores across all problem areas of ASEBA, while general community settings consist largely of healthy individuals and will therefore contribute lower scores overall.

Informant. The type of informant was coded based on which form of ASEBA was used (CBCL vs. YSR. vs. TRF) (caregiver, youth self-report, or teacher report). In the meta-regressions, two dummy codes were used, with caregiver as the comparison group.

Quality of Design and Reporting. Quality was calculated as a percentage of points earned out of total points possible on the quality items. We expected that higher quality designs would actually have lower overall effect sizes, which may seem counterintuitive. The quality ratings give more points to studies who do not utilize distilled designs or overfit thresholds (both of which artificially inflate effect size). The quality was assessed through the use of two a priori measures. Firstly, we used a system created by Kowatch et al. (2005) to rate design features, including the utilization of multiple informants and a formal consensus process for diagnoses, using the accepted DSM criteria for the time and a reference standard that questions lifetime symptoms. Other criteria are the inclusion of additional diagnoses in the sample beyond the target diagnoses and to report whether comorbid diagnoses were made. Although these standards were created for studies of pediatric bipolar disorder, these design features are relevant and have been generalized to other recent studies evaluating diagnostic efficiency (Youngstrom et al., 2015). In addition to the Kowatch criteria for evaluating the study design, we included QUADAS-2 criteria which provides a framework for rating the overall quality of the reporting of results in the included publications (Whiting et al., 2011).

Measures Included

The Child Behavior Checklist (CBCL) is completed by the caregiver of the child, typically the mother or father, and the Teacher Report Form (TRF) is filled out by the youth's teacher(s) based on how he/she acts in school. The CBCL and TRF are used to assess children ages 6-18, while the Youth Self Report (YSR) is administered to teens aged 11-18 to assess their own behavior (Achenbach, 1991; Achenbach & Edelbrock, 1983; Achenbach & Rescorla, 2001). The three forms of ASEBA all consist of 118 questions regarding a wide range of behaviors and problems that occur in youths. Each item is scored on a 0-2 point Likert scale, with 0 = *not at all true*, 1 = *somewhat or sometimes true*, and 2 = *very true or often true*. There are slight variations between the items on each form of ASEBA depending on the informant and their insight into the child's behavior. The scores are converted into a standardized *T* score for the child's age and sex.

A score is considered in the borderline clinical if it falls between 65-69, and is in the clinical range once it exceeds 70 (Achenbach, 1991).

Data Analytic Plan

In order to test the diagnostic efficiency of ASEBA, the sensitivity and specificity pair (or AUC) was converted to a Hedges' *g*, which is an effect size that corrects the small sample bias of Cohen's *d* (Hedges & Olkin, 1985). The use of AUC does not depend on a biased cut score or threshold as does the analysis of sensitivity and specificity, so it was therefore the better choice for our effect size estimate. All effect size estimates used a 95% confidence interval and inverse variance weighting to account for the size of both the sample and the effect size estimate in a way that gives the least weight to studies with the largest variance.

Many of the included studies contributed multiple effect sizes due to the use of multiple informants and/or the inclusion of sufficient information to calculate the diagnostic accuracy of more than one target diagnosis within the sample. This resulted in the nesting of effect sizes, which complicated our models because we could no longer assume that each effect size was independent from one another. *Metafor* (Viechtbauer, 2016b) is a statistical package in *R* (R Core Team, 2014) that allows the use of multivariate mixed meta-regression models (Berkey, Hoaglin, Antczak-Bouckoms, Mosteller, & Colditz, 1998; Gleser & Olkin, 2009) to test the hypotheses. This type of modeling shows the variance both within nested samples and between studies, which provided a direct comparison of how ASEBA performed in each study (rather than looking at a different model for each informant). Our model has three levels: random intercepts, a variable that models nesting within the sample (Konstantopoulous, 2011), and the sample weights in a block diagonal covariance matrix to show the dependence of nested effect sizes that were calculated using the same participants (Gleser & Olkin, 2009; Youngstrom et al., 2015; Youngstrom et al., 2018). At present, this is the most innovative and advanced method for performing this type of meta-analysis.

The effects of each potential moderator were tested individually and as well as in combination in a complete model. The predictors examined were: informant type, diagnostic category and definition, study setting, distilled sampling design, and ratings of quality. Cochran's Q tested homogeneity of the effect sizes – i.e., whether there was a significant difference in how ASEBA performed in various studies or if it worked about equally well in all instances. Funnel plots provided a visual indication of outliers as well as a publication bias, while Egger's test provided statistical analysis of publication bias. We also tested the power to make sure we could reject the null hypothesis of the effect size $g = 0$ (Hedges & Pigott, 2001).

Results

Figure 2 is a flow diagram displaying the search process. Our searches identified 28 articles published between 1994 and 2015 across 8 countries, contributing 223 unique effect sizes. The present study included 13,516 youths between the ages of 6 and 18 years. In terms of informant, 158 effect sizes came from caregiver report, 29 from teacher report, and 36 from youth self-report. All child and teacher effect sizes were nested within subsets of caregiver data. ADHD was the most frequently reported target diagnosis across studies, contributing 60 effect sizes; other target diagnoses identified include anxiety (*k*=41), bipolar spectrum (*k*=24), conduct disorder (*k*=16), depressive disorders (*k*=31), externalizing disorders (*k*=5), internalizing disorders (*k*=3), obsessive-compulsive disorder (*k*=1), oppositional defiant disorder (*k*=25), post-traumatic stress disorder (*k*=6), and substance use disorders (*k*=11). KSADS was the most common diagnostic interview contributing 131 effect sizes; both parent and youth were included in the diagnostic interview for 135 of the effect sizes. In terms of other moderators, 30% of studies used a distilled sampling design, and outpatient was the most common setting. Sixteen subscales of the ASEBA were reported across studies, with the most frequently reported being externalizing problems (*k*=44), attention problems (*k*=31), and internalizing problems (*k*=27). Table 1 provides a summary of sample-level characteristics.

The 223 effect sizes, moderator variables, and effect size level statistics are reported in Table 2. When multiple effect sizes were reported, we included all eligible estimates in the analyses. For example, individual studies often provided effect sizes for multiple target diagnoses, and different subscales were reported depending on each target diagnosis. As such, Table 2 reports the *N* for each effect size instead of providing a single estimate for the sample as a whole. Forest plots display the raw effect sizes. Given the number of effect sizes, figures break out separate forest plots for each target diagnosis (see Figures 3-12). There is no forest plot for obsessive-compulsive disorder, as there was only one effect size for this diagnosis.

Assessment of Study Quality

Study quality was assessed using two a priori measures. Kowatch criteria were used to assess design features of the included studies (e.g., multiple informants, consensus process for diagnosis, distilled sampling design, etc.) (Kowatch et al., 2005). When scaled as percent of the maximum possible score, the quality of design ranged from 34.62 to 92.31% with an average of 74.04%. The overall quality of the studies included was moderate in terms of using semi-structured interviews, implementing DSM criteria, capturing comorbid and confounding diagnoses, and other features that enhance confidence in the robustness of findings. The quality of reporting as defined by QUADAS-2 criteria was also moderate, with scores ranging from 28.95 to 94.74% with an average of 76.33%.

Overall Summary of Effect Sizes

Multivariate meta-regression (rma.mv in *metafor*) modeled the nesting of the 223 effect sizes in the 28 studies using random effects modeling for both the within study and between study variance estimates, simultaneously modeling the covariation between estimates based on the same participants. There was enormous heterogeneity of effect sizes, Cochran's $Q(222\ df) = 8712.81$, $p < .0001$. There were substantial variance components both for the within samples nesting of effect sizes (level 1 in a hierarchical linear model conceptual framework) – $\sigma^2 = .12$, as well as between samples (Level 2) – $\sigma^2 = .28$. This became the baseline model for further exploration of moderators and covariates. Table 3 reports the variance estimates and Cochran's *Q* for this model and subsequent augmented multivariate meta-regression models. The ICC of the true effects was large, $\rho = .74$. The average effect size, pooled across all measures and samples, was large: $g = 1.02$. The effect size indicates ASEBA generally shows good discriminative validity; however, the tremendous heterogeneity indicates there are differences too large to be solely attributed to sampling variation. This provides empirical motivation to test differences in the performance of ASEBA across informants, target diagnoses, and other potential moderators (Lipsey & Wilson, 2001).

Simple Moderator Analyses

Table 3 reports the variance estimates and Cochran's *Q* for the base model and the examination of each potential moderator individually. We first ran an augmented model examining the moderators individually and then created a fully saturated model to include all variables regardless of significance.

Informant. ASEBA school-age includes different versions for caregiver, teacher, and youth report in youths ages 6-18, allowing for information to be gathered from multiple informants. In the multivariate meta-regression, caregiver report was used as the reference category through the creation of dummy codes, to compare teacher versus caregiver and youth versus caregiver. This framework allowed the inclusion of all effect sizes and studies simultaneously (rather than running the analyses separately by informant). Type of informant explained a significant amount of heterogeneity, $Q(2\ df) = 13.83$, $p < .001$. Parameter estimates showed that caregiver report produced the largest effect size, $g = 1.06$, with teacher report averaging $g = -0.32$ lower and youth report $g = -0.27$ lower (all $p < .001$).

Target diagnosis. Diagnoses across studies were grouped, and included: ADHD, anxiety disorders, bipolar spectrum disorders, conduct disorder, depressive disorders, externalizing disorders, internalizing disorders, obsessive-compulsive disorder, oppositional defiant disorder, post-traumatic stress disorder, and substance use disorders. The most frequent target diagnosis across studies was ADHD, so we selected it as the reference for all other diagnoses to be evaluated against in the model. The effect size for ADHD (g = 0.96) was significantly lower than the average estimates for bipolar disorder (g = 0.24 higher), conduct disorder (g = 0.48 higher), and oppositional defiant disorder (g =0.31). Target diagnosis explained a significant amount of heterogeneity, Q(10 df) = 47.58, p < .0001.

Definition of target diagnosis. This moderator was dummy-coded to compare broad versus narrow definitions of the diagnosis. There were no significant differences in effect size whether the target diagnosis was defined narrowly or broadly, Q(1 df) = 1.68, p > .05.

Setting. The studies were grouped based on the settings from which participants were gathered: general community, at-risk individuals, general outpatient clinics, specialty outpatient centers, or inpatient units. In the simple model before adjusting for any other covariates, there were significant differences in effect sizes by setting. Outpatient setting produced significantly higher g estimates, B = 1.39, Q(4 df) = 13.77, p = .0081.

Distilled sample design. This variable was dummy-coded to compare the impact of distilled sampling design to samples with clinically-generalizable designs. There were no significant differences in effect size by sampling design, Q(1 df) = 0.99, p > .05.

Study quality. There were no significant differences in effect size based on quality of reporting as indexed by QUADAS-2 total, Q(1 df) = 0.01, p > .05. Likewise, the quality of study design as indexed by Kowatch scale total was not associated with effect size, $Q(1\ df) = 0.02$, $p > .05$.

Main Model: Multivariate Meta-Regression with All Predictors

A fully augmented model included all the moderators of interest simultaneously. This model accounted for substantial variance, $Q(33\ df)$ 135.43, $p < .0001$. It also reduced the random effect variance components both at Level 1 (within samples) – $\sigma^2 = .073$ versus $\sigma^2 = .12$ for the model with no moderators (a 39% reduction in heterogeneity), as well as Level 2 (between samples) – $\sigma^2 = .12$ versus $\sigma^2 = .28$ in the initial model, a 58% reduction at the between samples level. There still was significant remaining heterogeneity, Cochran's $Q(189\ df) = 3536.17$, $p < .0001$. Likelihood plot profiles suggested that the model provided accurate estimates, and the intraclass correlation between the estimated and true effects was 0.39 (Konstantopoulos, 2011).

Table 4 presents the regression weights and confidence intervals for the fully augmented model. The intercept is $b = 1.17$, $p < .001$, meaning that the average effect size is $g \sim 1$ for caregiver report of a child diagnosed with ADHD in an outpatient setting with a clinically generalizable (non-distilled) design. Informant, target diagnosis, and setting remained significant in the full model; diagnostic definition, distilled design, and quality remained insignificant.

Testing the Robustness of the Meta-Regression Models

Outlier analysis. Examination of standardized residuals flagged seven studies as potential outliers in the multivariate analyses; each reported an effect size more than 1.0 *g* larger than would be predicted based on the meta-regression model, $p<.005$, with standardized residuals $z > 2.5$. Rerunning the model with those seven studies excluded did not change the substantive pattern of findings; all moderators remained significant.

Publication bias. The analyses described above checked for influential outliers and examined the effects of omitting outliers on sensitivity analyses. We also used a multivariate extension of Egger's test by including weights of effect sizes in the model. This provides an estimate of symmetry in the data, where asymmetricity would indicate possible publication bias. We found no evidence of publication bias in the fully augmented model, $p = .16$.

Clinical Interpretability

In order to provide more clinically meaningful descriptions of the results, we saved the predicted values and confidence intervals from the meta-regression. We estimated the predicted values for each diagnosis, sorted by informant (see Table 5). We chose reference values that are likely to reflect the common possible implementation; this included using an outpatient setting, with broadly defined conditions, in a non-distilled sample. We also continued to include quality in the model, via the 75% percentile scores on QUADAS-2 and Kowatch. These predictions may change slightly in other settings and conditions; therefore, we do not have an empirical basis for making recommendations based on these characteristics.

Discussion

ASEBA school-age is a widely-used collection of instruments for assessing problem behaviors in children ages 6-18. The goal of the present study was to conduct a systematic review and meta-analyze the studies that reported the diagnostic accuracy of ASEBA to distinguish youth with a target diagnosis from other youths in the sample. This study also tested hypothesized moderators affecting the performance of ASEBA in published articles, including informant, diagnosis of interest, and study characteristics such as setting, use of distilled sampling methods, and quality of reporting and study design. The average effect sizes across all studies was large, indicating ASEBA shows good discriminative validity overall. The included studies often reported multiple effect sizes nested within the same sample, which was modeled through multivariate meta-regression in our analyses. The differences in studies at the sample level contributed to a considerable random effect variance component, and this was consistently larger than the variance component due to within-sample variation. The between-studies variance decreased for the fully augmented model including all moderators of interest; however, significant between-study variation still remained.

Moderators of Diagnostic Accuracy Effect Sizes

Informant effects. Caregiver report consistently outperformed teacher or youth report, showing larger effect sizes across each model included in the analyses. When all other moderators were conrolled for, the gap between caregiver and teacher report was g = -0.32 lower, and while youth report was g = -0.27 less than caregiver report. Teacher and youth report performed similarly, as shown by CIs that overlapped almost entirely. Caregiver report frequently yielded larger effect sizes than teacher or youth report in samples with nested effect sizes. These findings are consistent with other studies showing the superior accuracy of caregiver report over teacher or youth report (Youngstrom et al., 2004; 2005; 2015).

Closer examination of the raw effect sizes in table 2 shows the effect sizes ranged from $g = -0.61$ to 2.66 for caregiver report, from $g = -0.44$ to 1.36 for teacher report, and $g = -0.20$ to 1.06 for youth self-report. Once transformed into predicted values based on ideal conditions (outpatient setting with clinically realistic (non-distilled) sampling with broad diagnostic definition), the effect sizes ranged from $g = 0.92$ to 1.76 for caregiver report, $g = 0.14$ to 2.04 for teacher report, and $g = 0.53$ to 1.44 for youth report. As such, the large effect sizes seen for caregiver report indicate high discriminative ability regardless of target diagnosis, and the medium-large effect sizes for youth report indicate moderate-high discriminative validity. The spread of predicted effect sizes for teacher report was large; it corresponds with both the highest predicted effect size and the lowest predicted effect size, demonstrating the discriminative validity of teacher report varies widely depending on target diagnosis.

Target Diagnoses. Included studies evaluated the diagnostic accuracy of ASEBA for a range of diagnoses of interest, including: attention-deficit hyperactivity disorder, anxiety disorders, bipolar spectrum disorders, conduct disorder, depressive disorders, externalizing disorders, internalizing disorders, obsessive-compulsive disorder, oppositional defiant disorder, post-traumatic stress disorder, and substance use disorders. Target diagnosis was significant both when examined individually and when included in the fully augmented model, indicating type of diagnosis explained differences in how ASEBA performed in our model. ASEBA yielded significantly higher effect sizes for bipolar disorder, conduct disorder, and oppostional defiant disorder than ADHD, indicating it does a better job discriminating these conditions compared to

ADHD. The observed effect sizes for each disorder were widely spread, due to a variety of reasons (e.g., informant, setting, definition of diagnosis, comparison group characteristics, distilled design), which obscures the findings. In order to translate our findings to be clinically useful, we used the predicted values in Table 5 to infer ASEBA performance by disorder in the ideal conditions mentioned previously. The largest predicted effect size was associated with OCD (g = 1.76 for caregiver, 2.04 for teacher, and 1.44 for youth), indicating ASEBA shows high discriminative validity for OCD regardless of informant. However, this result should be interpreted with caution as OCD only contributed one of 223 effect sizes in the total analyses. Although each target diagnosis was associated with a range of predicted effect sizes based on informant, the effect sizes were large and showed high discriminative validity across all diagnoses.

Definition of diagnosis. We were interested in exploring whether using a broad or narrow diagnosis would impact the discriminative validity of ASEBA. There were no differences found in the effect sizes of broad versus narrow definitions, both when the variable was examined alone and when it was included in the full model with other predictors. We decided to include this moderator in the full model despite not showing significance when examined in isolation because it was of conceptual interest to us. Broad definition would include individuals with spectrum, subthreshold, and/or comorbid disorders, which is more consistent with what we would see in practice.

Setting. ASEBA is widely used to detect problem behaviors in children across settings, including schools, research studies, outpatient clinics, and inpatient hospitals. The present study showed significant differences in how the ASEBA performed in various settings. The largest effect sizes came from outpatient settings, indicating ASEBA performed best in the samples including this setting. Interestingly, the standardization sample of ASEBA is most similar to a general community setting. The assessments were normed using a sample of non-referred children which excluded any individuals who had connected with mental health or substance abuse services within the prior year. The recommended cut-offs are used across settings, regardless of whether the setting matches the characteristics of the normative sample on which the thresholds are based. In the outpatient studies in the present analyses, ASEBA was exclusively used for discriminating a target condition from other psychiatric disorders. The finding that ASEBA performed better in outpatient settings that utilized control groups that included multiple diagnoses, compared to the community settings which included healthy individuals, is counterintuitive. However, it provides support for the continued use of ASEBA as a screening measure in outpatient clinics.

Distilled sample enrollment. The inclusion of healthy controls can cause an artificial inflation of effect size due to the fact that healthy individuals score low on measures. Those with any disorder will score higher than the healthy controls, which makes the gap between disorders smaller than the gap between one disorder and heatlhy controls. This is important because the inflated effect sizes of a distilled sample will make it appear as if the measure is performing better in that study (Youngstrom et al., 2015; 2018). This superior performance may be a result of the sampling design and not the measure itself. As such, attention should be given to whether a sample utilizes a distilled design, and caution should be used when generalizing the findings of these studies. However, no significant differences were found in the effect sizes of distilled versus nondistilled designs in the present analysis, which could be due to the fact that only a third of the included studies utilized a distilled design. We chose to include this variable in the full model regardless due to its theoretical importance and significance in other recent meta-analyses examining diagnostic efficiency (Youngsrom et al., 2015; 2018).

Quality of publication. The Kowatch quality rating scale takes into consideration additional design features that could impact the effect sizes, for example the utilization of multiple informants and a consensus process when making diagnoses. Studies with a higher Kowatch score could be expected to yield smaller effect sizes because the studies would not be designed in ways that would cause artifical inflation (such as distilled samples, as discussed above). Additionally, we were interested in whether there was an impact of the quality of the reporting as indexed by the QUADAS-2. This study found no significant differences in effect sizes based on quality as defined by Kowatch and QUADAS-2 criteria. In addition to the present study, quality was not found to be significant in two other meta-analyses that assessed diagnostic efficiency (Youngstrom et al., 2015; 2018). We decided to leave this moderator in our analysis regardless because we believe quality study design and reporting is imporant. Additionally, many of the included studies were published before standardized guidelines for reporting were published (e.g.,STARD; Bossuyt et al., 2003b). Quality of reporting was not associated with effect sizes, which means results were not biased by quality.

Consideration of Alternative Explanations

One potential bias that may effect meta-analyses is called the "file drawer problem," which is concerned over whether the published literature is different from results that were not published. To assess whether this was a concern for our analysis we performed Egger's test, which found no evidence of publication bias. Another concern for the present study is the possible criterion contamination resulting from the use of caregiver informant in both the diagonstic interview and the index test. If diagnoses are based on information given by the caregiver, it conceptually makes sense for the index test score to be highly correlated with the results of the reference standard. Every study in the present analysis reported caregiver report (CBCL), and 71% of the total effect sizes are based on caregiver report. Additionally, 50% of the studies only interviewed caregiver when conducting the diagnostic interview. It is possible that the larger observed effect sizes of caregiver report may be due to the overlap in informant for ASEBA and the reference standard.

Generalizability of Conclusions

The studies included in this analyses examined the diagnostic accuracy of ASEBA for 11 target diagnoses in 13,516 youth across eight countries. The majority of our effect sizes came from the United States (65%), examined ADHD as the target diagnosis (38%), and took place in an outpatient setting (42%). As a result, the findings of our study show good generalizability for use of ASEBA in the United States under similar conditions. We are less confident in the generalizability of these findings in cases with fewer effect sizes available to analyze (e.g., OCD contributed 1 of 223 effect sizes; only 2 effect sizes came from Brazil). More data would be helpful to inform the generalizability in these instances. As mentioned earlier, the generalizability of the predicted effect sizes discussed in this paper depends on how similar or different the applied setting will be compared to the reference circumstances we used in the calculations.

Limitations

A general limitation of meta-analyses is the data included depends on the quality of the included studies. As discussed throughout the paper, effect sizes can be biased by various factors including distilled sampling, criterion contamination from informant, etc. As a result, our analyses could be biased due to any biases inherent in the samples included in the present study.

Additionally, although we found significant effects for several of our moderators of interest (informant, setting, and target diagnosis), substantial heterogeneity remains both between and within studies. This indicates the influence of moderators not included in the present analysis; the identification and assessment of these additional moderators is an important future direction.

Future Directions

The findings of the present study provide support for the continued use of ASEBA for detecting common behavior problems in youth. Through the course of this meta-analysis, we identified additional variables that were outside the scope of our present study but may yield many interesting results. Of note, interaction effects between our moderators of interest will provide a deeper understanding of the circumstances that may affect the relative performance of ASEBA (e.g., the relative accuracy of various subscales for a target disorder). In instances when ASEBA is found to be less accurate (i.e., certain diagnoses), another specialized measure could be added to address content gaps (Youngstrom et al., 2014). As noted in the limitations section, it will be imperative to explore additional moderators to address the significant heterogeneity that remains in our model.

ASEBA was used across many settings, although only 14% of the effect sizes came from a general community sample similar to the standardization sample. It may be helpful to re-norm ASEBA and update thresholds based on a sample more consistent with how the measure is being used. Another future direction could be the comparison of ASEBA with other broadband checklists for behavior problems in youth. If ASEBA is outperformed by the other checklist, it could indicate the replacement of ASEBA with the better-performing measure, especially if it has a lower cost of administration or is easier to use.

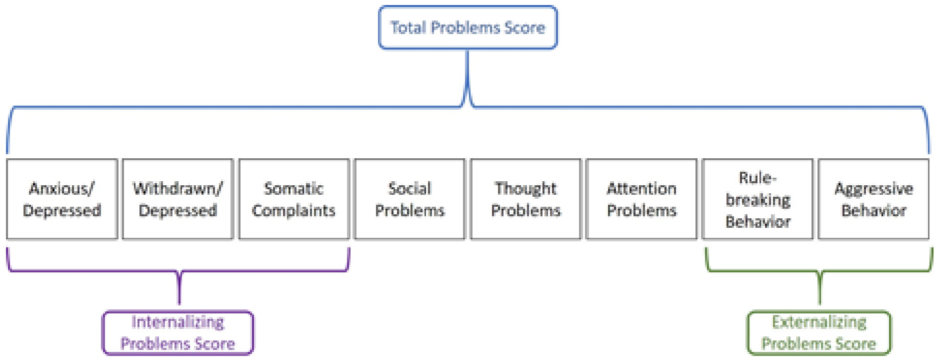

Figure 1. Diagram of ASEBA Syndrome Scales

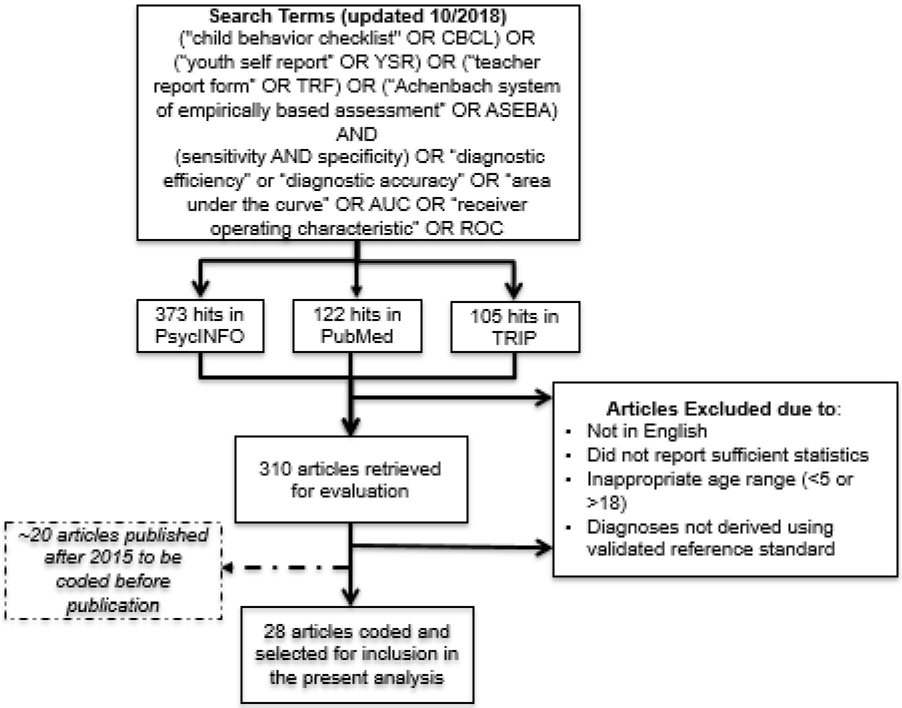

Figure 2. Flow diagram of included studies.

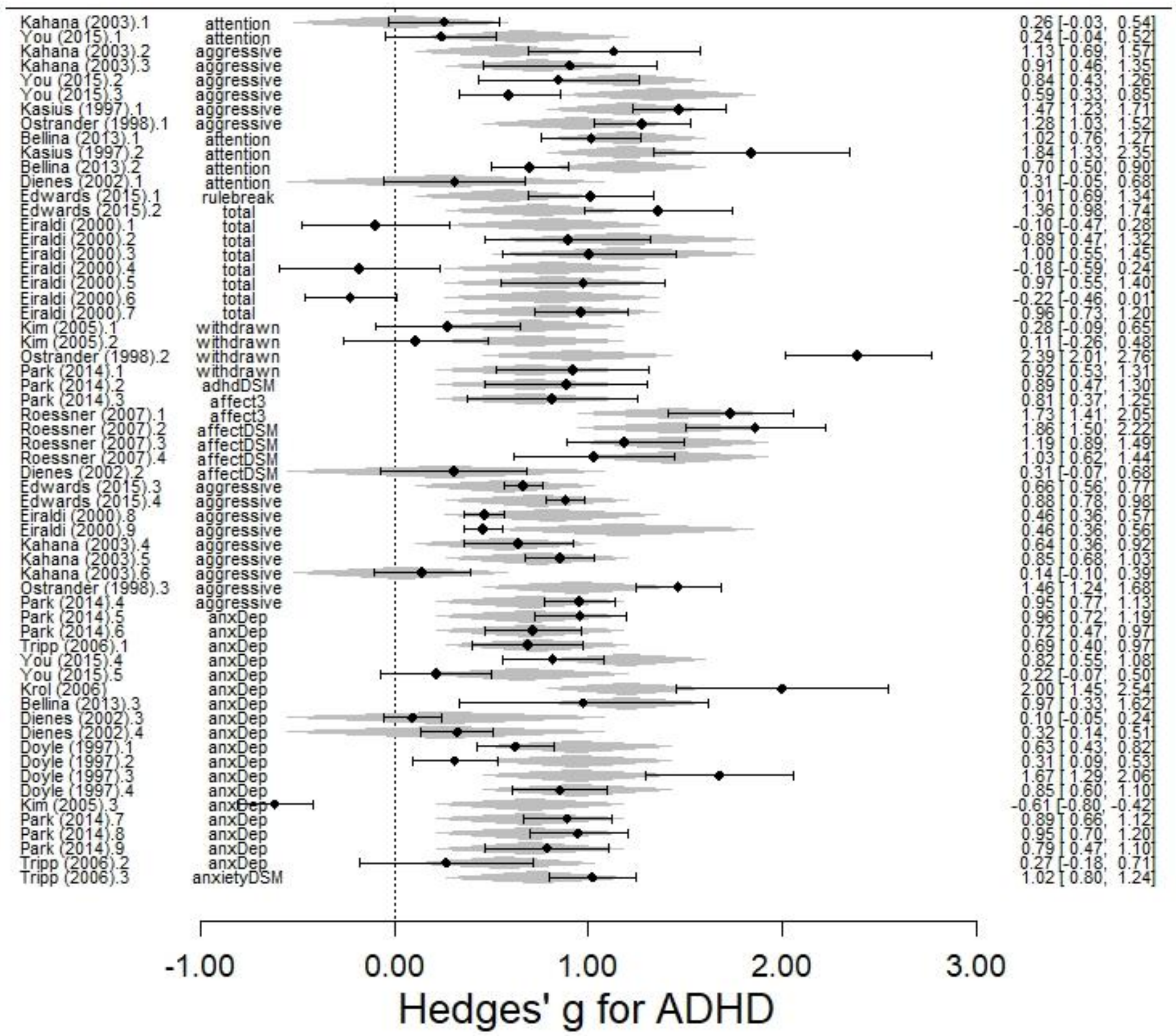

Figure 3. Forest plot for ADHD.

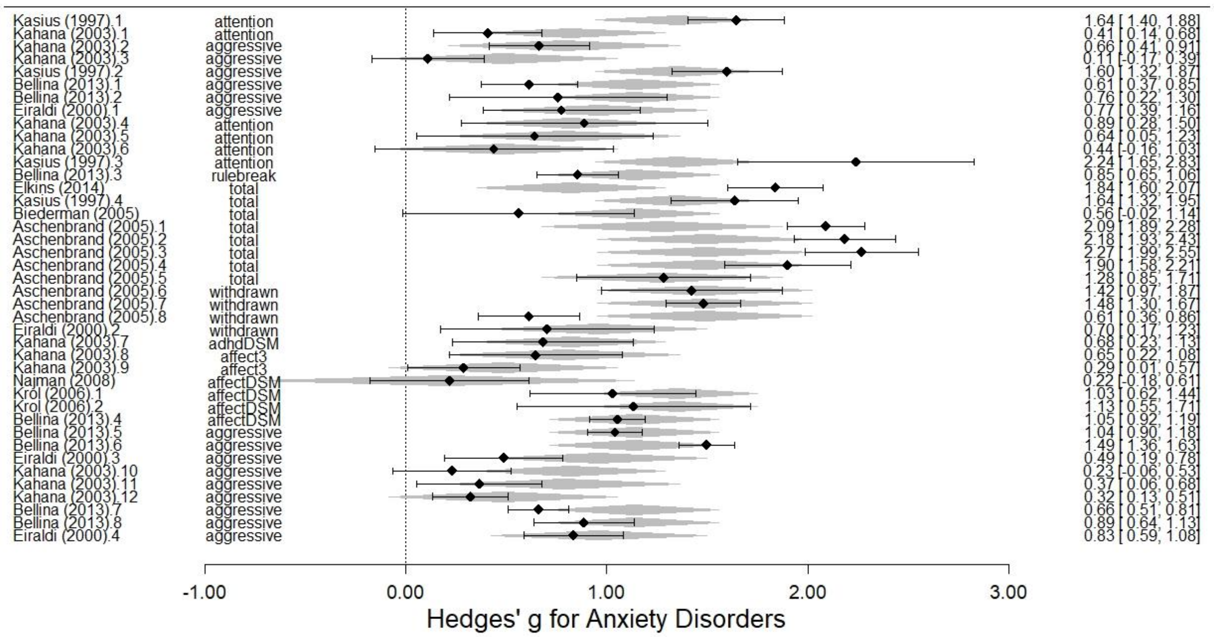

Figure 4. Forest plot for anxiety disorders.

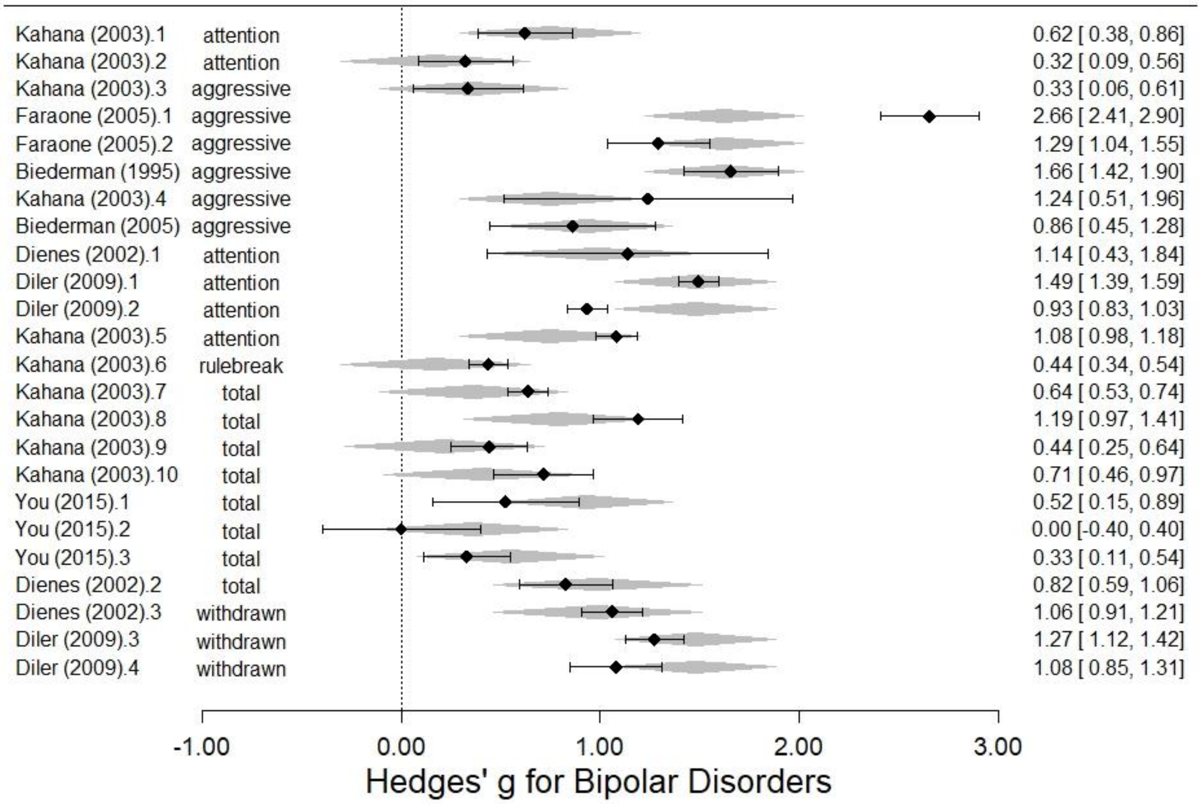

Figure 5. Forest plot for bipolar spectrum disorders.

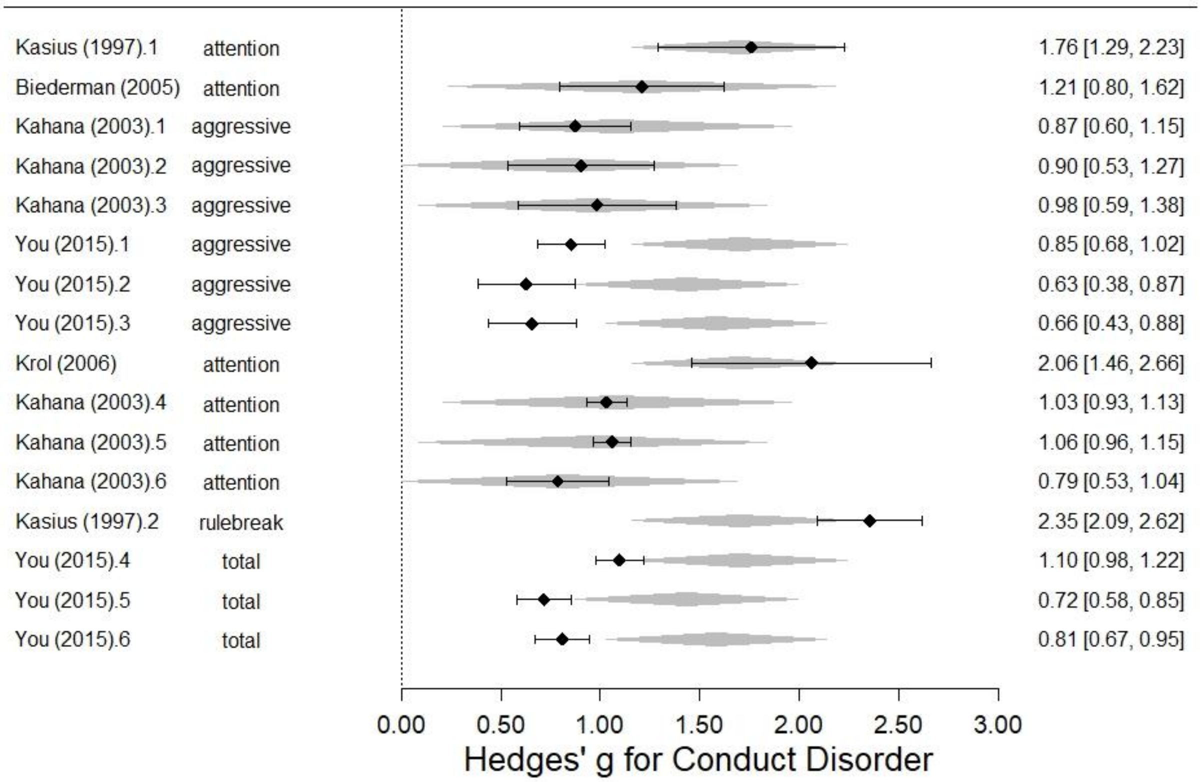

Figure 6. Forest plot for conduct disorder.

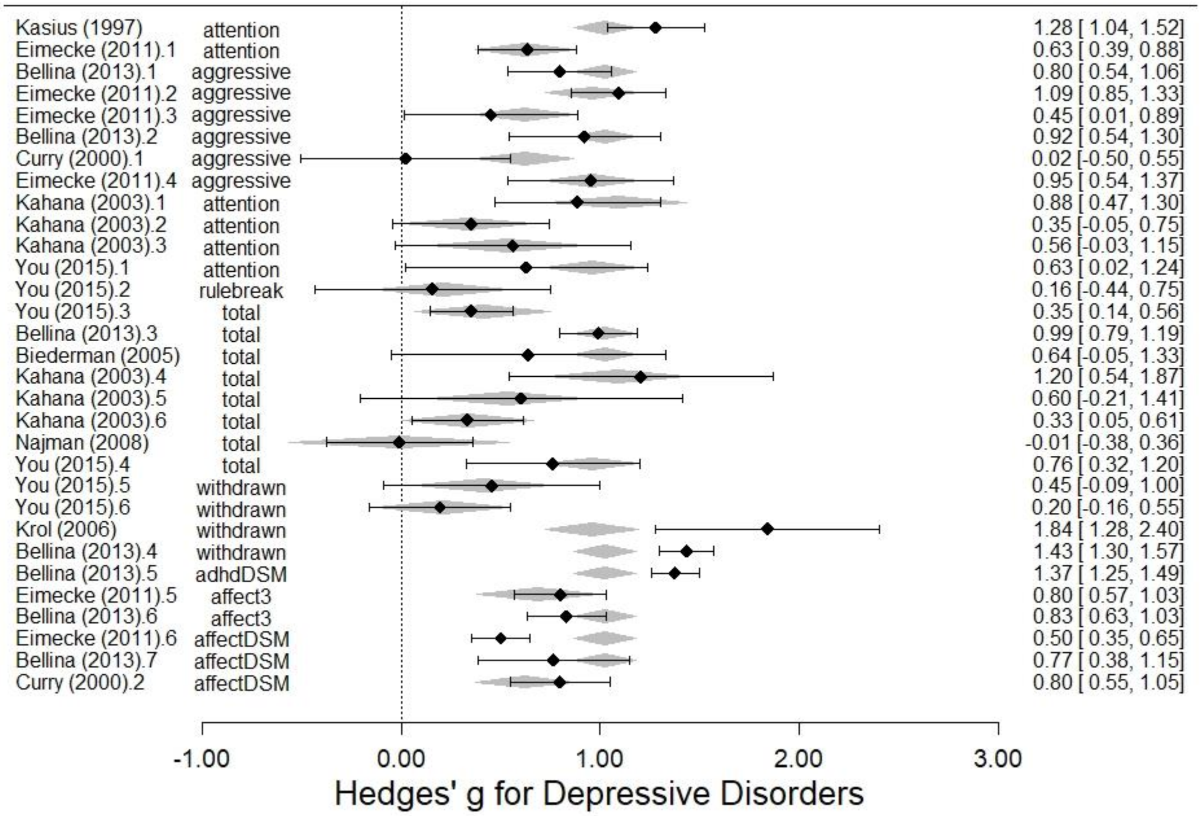

Figure 7. Forest plot for depressive disorders.

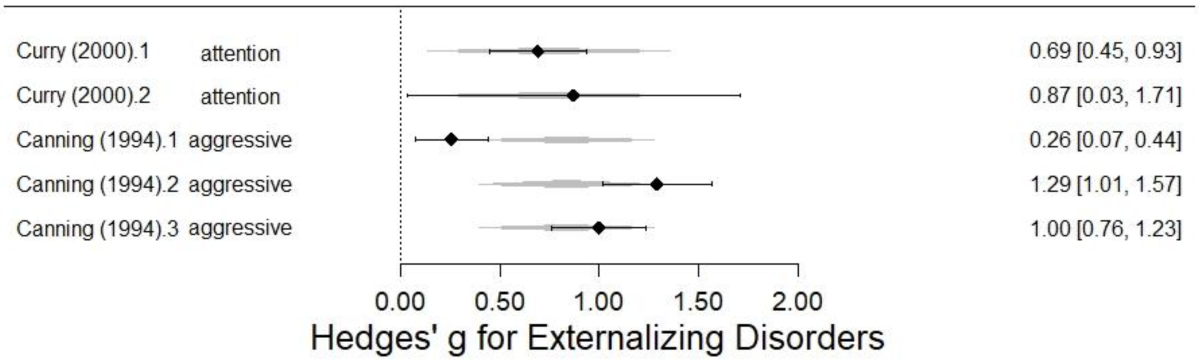

Figure 8. Forest plot for externalizing disorders.

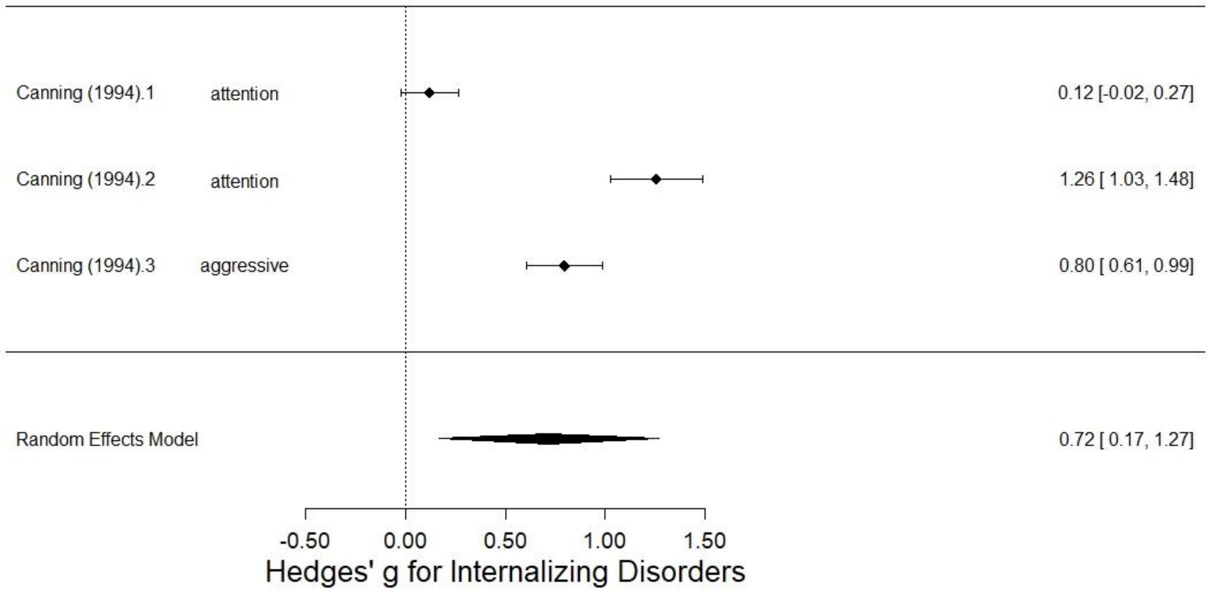

Figure 9. Forest plot for internalizing disorders.

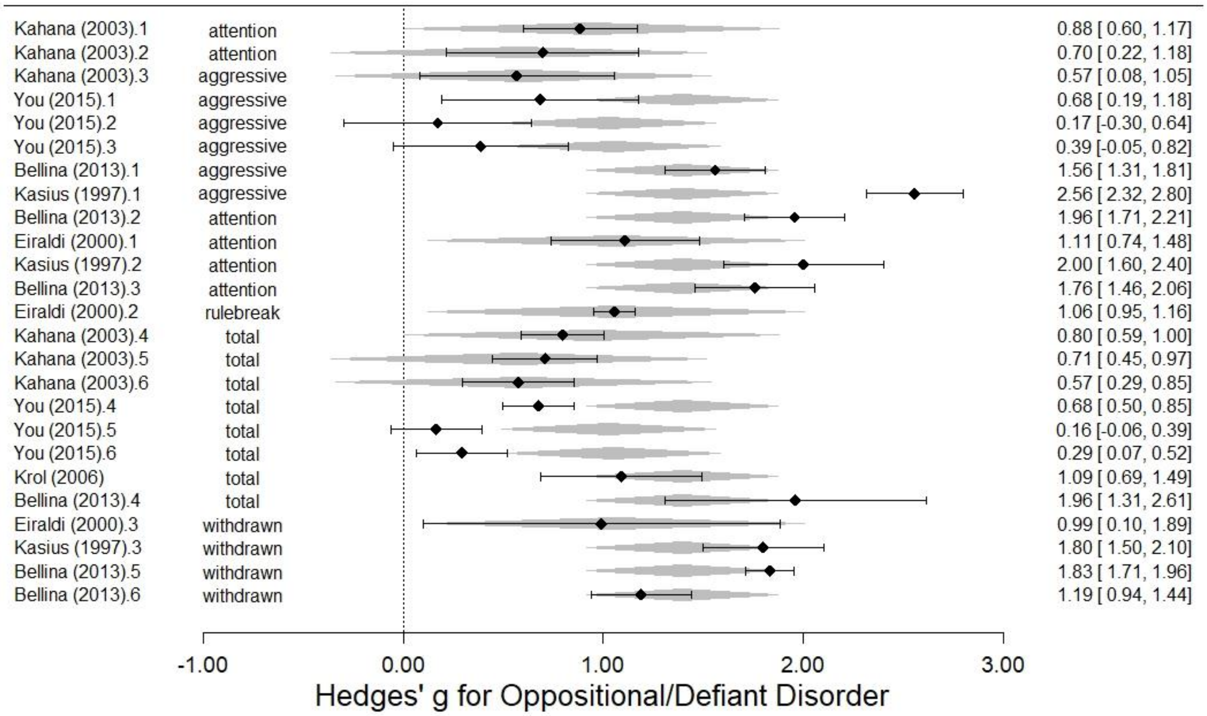

Figure 10. Forest plot for oppositional defiant disorder.

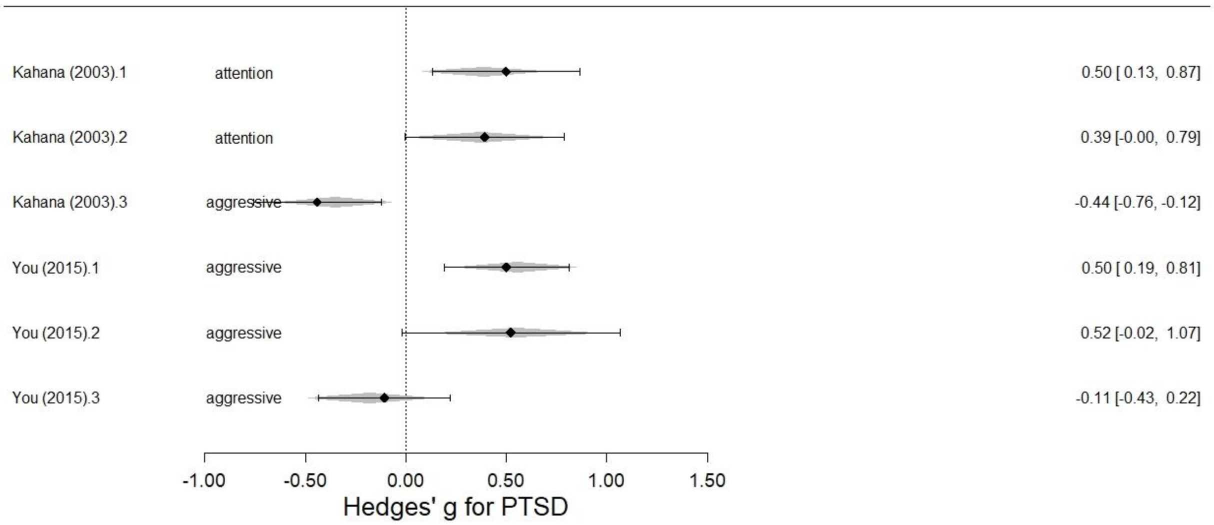

Figure 11. Forest plot for post-traumatic stress disorder.

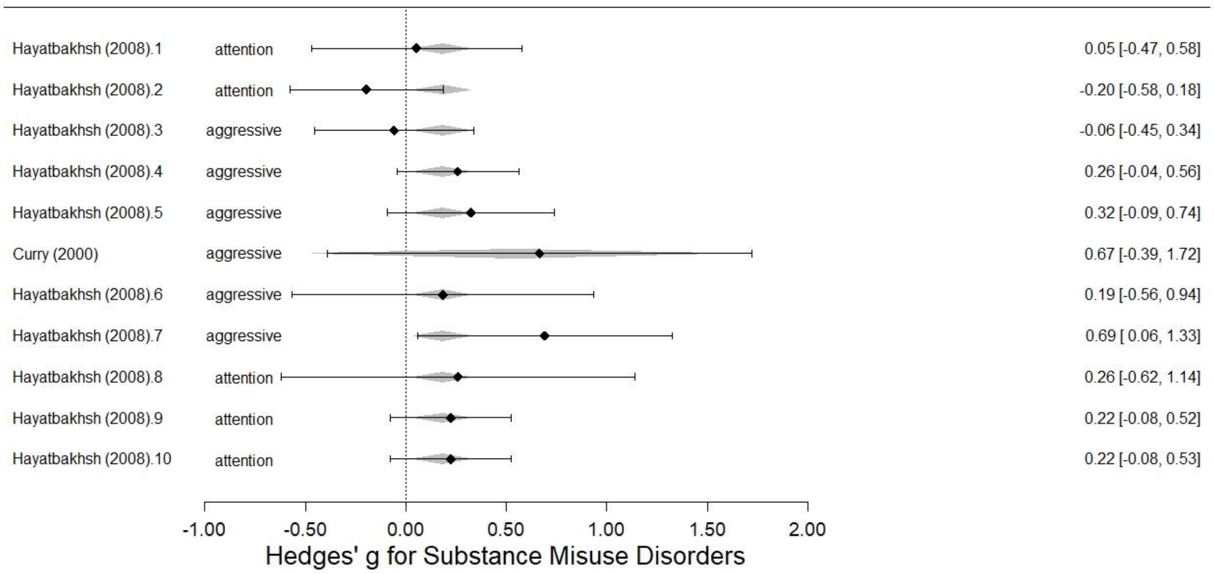

Figure 12. Forest plot for substance misuse disorders.

	Study level characteristics				
Study	Nested effects	Country	Setting type	Informant	Diagnostic Interview
Aschenbrand (2005)	8	USA	Outpatient	Caregiver	ADIS
Bellina (2013)	24	Italy	Outpatient	Caregiver	KSADS
Biederman (1995)	1	USA	Outpatient	Caregiver	KSADS
Biederman (2005)	4	USA	Outpatient	Caregiver	KSADS
Canning (1994)	6	USA	Specialty	Caregiver	DISC
Curry (2000)	5	USA	Inpatient	Caregiver	KSADS
Dienes (2002)	7	USA	At risk	Caregiver	KSADS
Diler (2009)	4	USA	Specialty	Caregiver	KSADS
Doyle (1997)	4	USA	Community	Caregiver	DICA
Edwards (2015)	4	USA	Specialty	Caregiver, Teacher	DISC
Eimecke (2011)	6	Germany	Outpatient + Inpatient	Caregiver	MAS
Eiraldi (2000)	16	USA	Specialty	Caregiver	DICA
Elkins (2014)	1	USA	Specialty	Caregiver	ADIS
Faraone (2005)	2	USA	Outpatient	Caregiver	KSADS
Geller (2006)	1	USA	Outpatient	Caregiver	KSADS
Hayatbakhsh (2008)	10	Australia	Community	Caregiver, Self	CIDI
Kahana (2003)	49	USA	Specialty	Caregiver, Teacher, Self	KSADS
Kasius (1997)	12	Netherlands	Outpatient	Caregiver	DISC
Kim (2005)	3	South Korea	Community	Caregiver	KSADS
Krol (2006)	6	Netherlands	Outpatient	Caregiver	DISC
Najman (2008)	2	Australia	Community	Caregiver	CIDI
Ostrander (1998)	3	USA	Community	Caregiver	DICA
Park (2014)	9	South Korea	Community	Caregiver	DISC
Roessner (2007)	4	Brazil + Germany	Outpatient	Caregiver	KSADS
Tripp (2006)	3	New Zealand	Specialty	Caregiver, Teacher	ADIS
You (2015)	29	USA	Outpatient	Caregiver, Teacher, Self	KSADS

Table 1. Summary of Sample-Level Characteristics of Studies Included in Meta-Analysis.

							Effect size	
Study	Nested Effects	*N* Target Diagnosis	*N* Comparison	Scale	Informant	Target Diagnosis	Hedge's g	Weight
Aschenbrand (2005)	8	100	30	INT	Caregiver	Anxiety	2.09	0.06
		100	30	INT	Caregiver	Anxiety	2.18	0.06
		100	30	INT	Caregiver	Anxiety	2.27	0.06
		100	30	INT	Caregiver	Anxiety	1.90	0.06
		100	30	INT	Caregiver	Anxiety	1.28	0.05
		100	30	INT	Caregiver	Anxiety	1.42	0.05
		100	30	INT	Caregiver	Anxiety	1.48	0.05
		100	30	INT	Caregiver	Anxiety	0.61	0.04
Bellina (2013)	24	125	173	Rule-break	Caregiver	ODD	1.56	0.02
		195	103	DSM-ADHD	Caregiver	ADHD	1.02	0.02
		125	173	DSM-AFF	Caregiver	Depression	0.80	0.01
		178	120	DSM-AFF	Caregiver	Anxiety	0.61	0.01
		125	173	Aggressive	Caregiver	ODD	1.96	0.02
		125	173	Anx/dep	Caregiver	Depression	0.92	0.02
		178	120	Anx/dep	Caregiver	Anxiety	0.76	0.01
		125	173	DSM-ANX	Caregiver	Depression	0.99	0.02
		178	120	DSM-ANX	Caregiver	Anxiety	0.85	0.02
		195	103	Attention	Caregiver	ADHD	0.70	0.02
		125	173	DSM-CD	Caregiver	ODD	1.76	0.02
		195	103	DSM-ODD	Caregiver	ADHD	0.97	0.02
		125	173	DSM-ODD	Caregiver	ODD	1.96	0.02
		125	173	Social	Caregiver	ODD	1.83	0.02
		125	173	Social	Caregiver	Depression	1.43	0.02
		178	120	Social	Caregiver	Anxiety	1.05	0.02
		125	173	Somatic	Caregiver	Depression	1.37	0.02
		178	120	Somatic	Caregiver	Anxiety	1.04	0.02
		178	120	Somatic	Caregiver	Anxiety	1.49	0.02
		125	173	Total	Caregiver	Depression	0.83	0.01
		178	120	Total	Caregiver	Anxiety	0.66	0.01
		125	173	Total	Caregiver	ODD	1.19	0.02
		125	173	Withdrawn	Caregiver	Depression	0.77	0.01
		178	120	Withdrawn	Caregiver	Anxiety	0.89	0.02
Biederman (1995)	1	43	164	Aggressive	Caregiver	Bipolar	1.66	0.04

Biederman (2005)	4	18	103	EXT	Caregiver	CD	1.21	0.07
		8	113	EXT	Caregiver	Bipolar	0.86	0.14
		18	103	EXT	Caregiver	Depression	0.64	0.07
		35	86	EXT	Caregiver	Anxiety	0.56	0.04
Canning (1994)	6	21	100	Total	Caregiver	Internalizing disorder	0.12	0.06
		31	90	Total	Caregiver	Internalizing disorder	1.26	0.05
		45	76	Total	Caregiver	Internalizing disorder	0.80	0.04
		15	106	Total	Caregiver	Externalizing disorder	0.26	0.08
		18	103	Total	Caregiver	Externalizing disorder	1.29	0.07
		34	87	Total	Caregiver	Externalizing disorder	1.00	0.05
Curry (2000)	5	57	51	Aggressive	Caregiver	Externalizing disorder	0.69	0.04
		46	62	Anx/dep	Caregiver	Depression	0.02	0.04
		57	51	Rule-break	Caregiver	Externalizing disorder	0.87	0.04
		34	74	Rule-break	Caregiver	Substance use disorder	0.67	0.04
		46	62	Withdrawn	Caregiver	Depression	0.80	0.04
Dienes (2002)	7	15	43	Attention	Caregiver	ADHD	0.31	0.09
		16	42	EXT	Caregiver	Bipolar	1.14	0.10
		15	43	EXT	Caregiver	ADHD	0.31	0.09
		16	42	INT	Caregiver	Bipolar	0.82	0.09
		15	43	Thought	Caregiver	ADHD	0.10	0.09
		16	42	Total	Caregiver	Bipolar	1.06	0.10
		15	43	Total	Caregiver	ADHD	0.32	0.09
Diler (2009)	4	157	356	EXT	Caregiver	Bipolar	1.49	0.01
		157	356	EXT	Caregiver	Bipolar	0.93	0.01
		157	356	Total	Caregiver	Bipolar	1.27	0.01
		157	356	Total	Caregiver	Bipolar	1.08	0.01
Doyle (1997)	4	115	41	Total	Caregiver	ADHD	0.63	0.03
		60	96	Total	Caregiver	ADHD	0.31	0.03

		55	101	Total	Caregiver	ADHD	1.67	0.04
		60	55	Total	Caregiver	ADHD	0.85	0.04
Edwards (2015)	4	46	49	Attention	Caregiver	ADHD	1.01	0.05
		46	49	Attention	Teacher	ADHD	1.36	0.05
		46	49	EXT	Caregiver	ADHD	0.66	0.04
		46	49	EXT	Teacher	ADHD	0.88	0.05
Eimecke (2011) -a	6	53	540	DSM-AFF	Caregiver	Depression	0.63	0.02
Eimecke (2011) -a		53	540	Anx/dep	Caregiver	Depression	0.45	0.02
Eimecke (2011) -a		53	540	Total	Caregiver	Depression	0.80	0.02
Eimecke (2011)-b		74	1282	Total	Caregiver	Depression	0.50	0.01
Eimecke (2011)-b		74	1282	DSM-AFF	Caregiver	Depression	1.09	0.01
Eimecke (2011)-b		74	1282	Anx/dep	Caregiver	Depression	0.95	0.01
Eiraldi (2000)	16	67	161	Aggressive	Caregiver	ODD	1.11	0.02
		26	202	Anx/dep	Caregiver	Anxiety	0.77	0.04
		115	58	Attention	Caregiver	ADHD	-0.10	0.03
		173	36	Attention	Caregiver	ADHD	0.89	0.04
		173	36	Attention	Caregiver	ADHD	1.00	0.04
		115	36	Attention	Caregiver	ADHD	-0.18	0.04
		115	36	Attention	Caregiver	ADHD	0.97	0.04
		58	36	Attention	Caregiver	ADHD	-0.22	0.05
		58	36	Attention	Caregiver	ADHD	0.96	0.05
		115	58	EXT	Caregiver	ADHD	0.46	0.03
		173	36	EXT	Caregiver	ADHD	0.46	0.03
		67	161	EXT	Caregiver	ODD	1.06	0.02
		26	202	INT	Caregiver	Anxiety	0.70	0.04
		67	161	Rule-break	Caregiver	ODD	0.99	0.02
		26	202	Somatic	Caregiver	Anxiety	0.49	0.04
		26	202	Withdrawn	Caregiver	Anxiety	0.83	0.04
Elkins (2014)	1	23	23	Attention	Caregiver	Anxiety	1.84	0.12
Faraone (2005)	2	13	458	PBD	Caregiver	Bipolar	2.66	0.09
		8	402	PBD	Caregiver	Bipolar	1.29	0.13

Geller (2006)	1	64	129	OCS	Caregiver	OCD	2.27	0.04
Hayatbakhsh (2008)	10	530	1397	Aggressive	Self	Substance use disorder	0.05	0.00
		530	1397	Anx/dep	Self	Substance use disorder	-0.20	0.00
		503	1397	Anx/dep	Self	Substance use disorder	-0.06	0.00
		530	1397	Attention	Self	Substance use disorder	0.26	0.00
		503	1397	Attention	Self	Substance use disorder	0.32	0.00
		530	1397	Rule-break	Self	Substance use disorder	0.19	0.00
		503	1397	Rule-break	Self	Substance use disorder	0.69	0.00
		503	1397	Rule-break	Self	Substance use disorder	0.26	0.00
		530	1397	Somatic	Self	Substance use disorder	0.22	0.00
		503	1397	Somatic	Self	Substance use disorder	0.22	0.00
Kahana (2003)	49	138	183	Attention	Self	ADHD	0.26	0.01
		148	424	Aggressive	Caregiver	ODD	0.88	0.01
		91	214	Aggressive	Teacher	ODD	0.70	0.02
		62	259	Aggressive	Self	ODD	0.57	0.02
		319	253	Attention	Caregiver	ADHD	1.13	0.01
		198	107	Attention	Teacher	ADHD	0.91	0.02
		157	236	Total	Caregiver	Bipolar	0.62	0.01
		98	135	Total	Teacher	Bipolar	0.32	0.02
		84	122	Total	Self	Bipolar	0.33	0.02

55	519	Withdrawn	Caregiver	Anxiety	0.41	0.02
32	290	Withdrawn	Self	Anxiety	0.66	0.04
27	279	Withdrawn	Teacher	Anxiety	0.11	0.04
157	236	Aggressive	Caregiver	Bipolar	1.24	0.01
383	196	Anx/dep	Caregiver	Depression	0.88	0.01
196	111	Anx/dep	Teacher	Depression	0.35	0.01
237	86	Anx/dep	Self	Depression	0.56	0.02
55	519	Anx/dep	Caregiver	Anxiety	0.89	0.02
32	290	Anx/dep	Self	Anxiety	0.64	0.04
27	279	Anx/dep	Teacher	Anxiety	0.44	0.04
157	236	EXT	Caregiver	Bipolar	1.08	0.01
98	135	EXT	Teacher	Bipolar	0.44	0.02
84	122	EXT	Self	Bipolar	0.64	0.02
263	309	EXT	Caregiver	Bipolar	1.19	0.01
151	154	EXT	Teacher	Bipolar	0.44	0.01
139	182	EXT	Self	Bipolar	0.71	0.01
319	253	EXT	Caregiver	ADHD	0.64	0.01
198	107	EXT	Teacher	ADHD	0.85	0.02
138	183	EXT	Self	ADHD	0.14	0.01
148	424	EXT	Caregiver	ODD	0.80	0.01
91	214	EXT	Teacher	ODD	0.71	0.02
62	259	EXT	Self	ODD	0.57	0.02
43	529	EXT	Caregiver	CD	0.87	0.03
23	282	EXT	Teacher	CD	0.90	0.05
21	300	EXT	Self	CD	0.98	0.05
383	196	INT	Caregiver	Depression	1.20	0.01
237	86	INT	Self	Depression	0.60	0.02
196	111	INT	Teacher	Depression	0.33	0.01
14	558	INT	Caregiver	PTSD	0.50	0.07
9	312	INT	Self	PTSD	0.39	0.11
6	299	INT	Teacher	PTSD	-0.44	0.17
55	519	INT	Caregiver	Anxiety	0.68	0.02
32	290	INT	Self	Anxiety	0.65	0.04
27	279	INT	Teacher	Anxiety	0.29	0.04
43	529	Rule-break	Caregiver	CD	1.03	0.03
21	300	Rule-break	Self	CD	1.06	0.05
23	282	Rule-break	Teacher	CD	0.79	0.05
55	518	Somatic	Caregiver	Anxiety	0.23	0.02
32	290	Somatic	Self	Anxiety	0.37	0.03
27	279	Somatic	Teacher	Anxiety	0.32	0.04

Kasius (1997)	12	33	198	Total	Caregiver	ODD	2.56	0.05
		15	216	Total	Caregiver	Anxiety	1.64	0.08
		42	189	Total	Caregiver	Depression	1.28	0.03
		70	161	Total	Caregiver	ADHD	1.47	0.03
		15	216	Withdrawn	Caregiver	Anxiety	1.60	0.08
		70	161	Aggressive	Caregiver	ADHD	1.84	0.03
		33	198	Aggressive	Caregiver	ODD	2.00	0.04
		13	218	Aggressive	Caregiver	CD	1.76	0.09
		15	216	Anx/dep	Caregiver	Anxiety	2.24	0.08
		15	216	Attention	Caregiver	Anxiety	1.64	0.08
		33	198	Rule-break	Caregiver	ODD	1.80	0.04
		13	218	Rule-break	Caregiver	CD	2.35	0.09
Kim (2005)	3	33	13	Attention	Caregiver	ADHD	0.28	0.11
		33	13	Attention	Caregiver	ADHD	0.11	0.11
		33	13	Total	Caregiver	ADHD	-0.61	0.11
Krol (2006)	6	7	37	DSM-ANX	Caregiver	Anxiety	1.03	0.18
		4	40	DSM-ANX	Caregiver	Anxiety	1.13	0.29
		7	37	DSM-DEP	Caregiver	Depression	1.84	0.21
		29	15	DSM-ADHD	Caregiver	ADHD	2.00	0.15
		22	22	DSM-ODD	Caregiver	ODD	1.09	0.10
		8	36	DSM-CD	Caregiver	CD	2.06	0.20
Najman (2008)	2	450	1861	INT	Caregiver	Depression	-0.01	0.00
		566	1747	INT	Caregiver	Anxiety	0.22	0.00
Ostrander (1998)	3	194	107	Total	Caregiver	ADHD	1.28	0.02
		194	107	Attention	Caregiver	ADHD	2.39	0.02
		194	107	EXT	Caregiver	ADHD	1.46	0.02
Park (2014)	9	481	740	Attention	Caregiver	ADHD	0.92	0.00
		379	582	Attention	Caregiver	ADHD	0.89	0.00
		356	547	Attention	Caregiver	ADHD	0.81	0.01
		481	740	EXT	Caregiver	ADHD	0.95	0.00
		379	582	EXT	Caregiver	ADHD	0.96	0.00
		356	547	EXT	Caregiver	ADHD	0.72	0.00
		481	740	Total	Caregiver	ADHD	0.89	0.00
		379	582	Total	Caregiver	ADHD	0.95	0.00
		356	547	Total	Caregiver	ADHD	0.79	0.00
Roessner (2007) -a	4	248	71	Attention	Caregiver	ADHD	1.73	0.02

Roessner (2007) -a		248	71	Attention	Caregiver	ADHD	1.86	0.02
Roessner (2007) -b		154	135	Attention	Caregiver	ADHD	1.19	0.02
Roessner (2007) -b		154	135	Attention	Caregiver	ADHD	1.03	0.02
Tripp (2006)	3	108	76	EXT	Teacher	ADHD	0.69	0.02
		108	76	Total	Caregiver	ADHD	0.27	0.02
		108	76	Total	Teacher	ADHD	1.02	0.03
You (2015)	29	231	212	Attention	Self	ADHD	0.24	0.01
		306	467	Aggressive	Caregiver	ODD	0.68	0.01
		128	164	Aggressive	Teacher	ODD	0.17	0.01
		155	285	Aggressive	Self	ODD	0.39	0.01
		493	283	Attention	Caregiver	ADHD	0.84	0.01
		199	94	Attention	Teacher	ADHD	0.59	0.02
		355	417	Anx/dep	Caregiver	Depression	0.63	0.01
		143	149	Anx/dep	Teacher	Depression	0.16	0.01
		251	188	Anx/dep	Self	Depression	0.35	0.01
		141	640	EXT	Caregiver	Bipolar	0.52	0.01
		65	229	EXT	Teacher	Bipolar	0.00	0.02
		86	362	EXT	Self	Bipolar	0.33	0.01
		493	283	EXT	Caregiver	ADHD	0.82	0.01
		231	212	EXT	Self	ADHD	0.22	0.01
		306	467	EXT	Caregiver	ODD	0.68	0.01
		128	164	EXT	Teacher	ODD	0.16	0.01
		155	285	EXT	Self	ODD	0.29	0.01
		94	681	EXT	Caregiver	CD	0.85	0.01
		30	263	EXT	Teacher	CD	0.63	0.04
		79	363	EXT	Self	CD	0.66	0.02
		355	417	INT	Caregiver	Depression	0.76	0.01
		251	188	INT	Self	Depression	0.45	0.01
		143	149	INT	Teacher	Depression	0.20	0.01
		67	704	INT	Caregiver	PTSD	0.50	0.02
		43	395	INT	Self	PTSD	0.52	0.03
		21	271	INT	Teacher	PTSD	-0.11	0.05
		94	681	Rule-break	Caregiver	CD	1.10	0.01
		30	263	Rule-break	Teacher	CD	0.72	0.04
		79	363	Rule-break	Self	CD	0.81	0.02

Note: INT = internalizing problems; EXT= externalizing problems; Rule-break = rule-breaking behavior subscale; DSM-ADHD = DSM-oriented ADHD scale; DSM-AFF = DSM-oriented affective disorder scale; Agg. = aggressive behavior subscale; Anx/Dep = anxious/depressed subscale; DSM-ANX = DSM-oriented anxiety scale; DSM-CD = DSM-oriented conduct problems scale; DSM-ODD = DSM-oriented oppositional defiance scale; Social = social problems subscale; Somatic

= somatic complaints; Withdrawn = withdrawn/depressed subscale; Attention = attention problems subscale; Thought = thought problems subscale; PBD = pediatric bipolar disorder profile; OCS = obsessive-compulsive scale.

Table 2. Effect Size Level Characteristics and Moderators.

Model	Level 1 variance	Level 2 Variance	*Q* Residual (*df*)	*Q* Model (*df*)
No moderators	0.122	0.276	8712.81 (222) ***	
Moderator: Informant	0.115	0.245	7760.80 (220) ***	13.83 (2) **
Moderator: Target Diagnosis	0.097	0.219	6632.71 (212) ***	47.58 (10) ***
Moderator: Diagnostic Definition	0.121	0.276	8510.77 (221)***	1.68 (1)
Moderator: Setting	0.115	0.146	7592.76 (216) ***	30.53 (6) ***
Moderator: Distilled Design	0.122	0.268	8529.45 (221)***	1.0 (1)
Moderator: QUADAS-2	0.122	0.275	8511.41 (221)***	0.01 (1)
Moderator: Kowatch	0.121	0.279	8571.75 (221)***	.02 (1)
Moderators: All simultaneously	0.073	0.115	3536.17 (189) ***	135.43 (33) ***

Table 3. Tests of Homogeneity and Estimates of Random Effects Variances Between Effect Sizes (Level 1) and Between Samples (Level 2) for Multivariate Meta-Regression Models Using Maximum Likelihood Estimation

Variable	*b*	*SE*	ci.lb	ci.ub
Intercept	1.17	0.19	0.81	1.53
Teacher report	0.28	0.14	.00	.56
Youth report	-0.32	0.18	-0.66	0.03
Anxiety	0.01	0.11	-0.21	0.23
Bipolar	0.37	0.13	0.11	0.64
Conduct	0.48	0.14	0.22	0.75
Depression	0.04	0.13	-0.21	0.29
Externalizing	0.17	0.30	-0.41	0.75
Internalizing	0.02	0.34	-0.65	0.69
OCD	0.78	0.54	-0.29	1.84
ODD	0.46	0.10	0.26	0.66
PTSD	-0.06	0.24	-0.53	0.41
SUD	0.07	0.62	-1.15	1.29
Narrow definition	-0.01	0.08	-0.17	0.15
Distilled sample	0.27	0.18	-0.08	0.62
Community setting	-0.46	0.27	-0.98	0.06
At-risk setting	-0.96	0.39	-1.72	-0.18
Specialty setting	-0.17	0.24	-0.64	0.31
Inpatient setting	-0.37	0.33	-1.02	0.29
QUADAS 75%	0.00	0.00	-0.02	0.02
Kowatch 75%	-0.35	0.22	-0.74	0.05

Table 4. Multivariate Meta-Regression Estimates of the Effects of Moderators Entered Together in the Model

Diagnostic Target	Informant	*g*	*SE*	ci.lb	ci.ub
ADHD	Teacher	1.26	0.78	-0.27	2.79
ADHD	Caregiver	0.98	0.76	-0.52	2.48
ADHD	Youth	0.66	0.77	-0.85	2.17
Anxiety Disorders	Caregiver	0.99	0.76	-0.50	2.48
Anxiety Disorders	Youth	0.93	0.77	-0.58	2.44
Anxiety Disorders	Teacher	0.65	0.77	-0.86	2.16
Bipolar Spectrum Disorders	Caregiver	1.35	0.76	-0.14	2.85
Bipolar Spectrum Disorders	Youth	0.92	0.77	-0.58	2.43
Bipolar Spectrum Disorders	Teacher	0.73	0.77	-0.78	2.23
Conduct Disorder	Caregiver	1.46	0.76	-0.03	2.95
Conduct Disorder	Youth	1.31	0.77	-0.19	2.82
Conduct Disorder	Teacher	1.20	0.77	-0.31	2.71
Depressive Disorders	Caregiver	1.02	0.75	-0.45	2.48
Depressive Disorders	Youth	0.89	0.77	-0.63	2.41
Depressive Disorders	Teacher	0.67	0.77	-0.84	2.17
Externalizing Disorders	Teacher	1.43	0.79	-0.11	2.98
Externalizing Disorders	Caregiver	1.15	0.77	-0.36	2.66
Externalizing Disorders	Youth	0.83	0.78	-0.70	2.36
Internalizing Disorders	Teacher	1.28	0.80	-0.29	2.85
Internalizing Disorders	Caregiver	1.00	0.78	-0.53	2.52
Internalizing Disorders	Youth	0.68	0.79	-0.88	2.23
OCD	Teacher	2.04	1.04	0.00	4.08
OCD	Caregiver	1.76	1.03	-0.26	3.77
OCD	Youth	1.44	1.03	-0.59	3.47
Oppositional Defiant Disorder	Caregiver	1.44	0.76	-0.05	2.93
Oppositional Defiant Disorder	Youth	0.86	0.77	-0.65	2.38
Oppositional Defiant Disorder	Teacher	0.84	0.77	-0.67	2.36
PTSD	Caregiver	0.92	0.79	-0.62	2.46
PTSD	Youth	0.85	0.79	-0.70	2.41
PTSD	Teacher	0.14	0.78	-1.40	1.68
Substance Use Disorders	Teacher	1.33	0.98	-0.59	3.25
Substance Use Disorders	Caregiver	1.05	0.96	-0.84	2.94
Substance Use Disorders	Youth	0.53	0.83	-1.10	2.17

Table 5. Summary of Estimated Effect Sizes by Target Diagnosis, Informant, and Sample Design

REFERENCES

Ablow, J. C., Measelle, J. R., Kraemer, H. C., Harrington, R., Luby, J., Smider, N., . . . Kupfer, D. J. (1999). The macarthur Three-City Outcome Study: Evaluating multi-informant measures of young children's symptomatology. *Journal of the American Academy of Child & Adolescent Psychiatry, 38*, 1580-1590.

Achenbach, T. M. (1991). *Manual for Child Behavior Checklist/ 4-18 and 1991 Profile.* Burlington, VT: University of Vermont, Dept. of Psychiatry.

Achenbach, T. M. (1995). Empirically based assessment and taxonomy: Applications to clinical research. *Psychological Assessment, 7*, 261-274. doi:10.1037/1040-3590.7.3.261

Achenbach, T. M. (2001). What are norms and why do we need valid ones? *Clinical Psychology: Science & Practice, 8*, 446-450. doi:10.1093/clipsy.8.4.446

Achenbach, T. M. (2009). *The Achenbach System of Empirically Based Assessemnt (ASEBA): Development, Findings, Theory, and Applications*. Burlington, VT: University of Vermont Research Center for Children, Youth, & Families.

Achenbach, T.M., Becker, A., Döpfner, M., Heiervang, E., Roessner, V., Steinhausen, H., & Rothenberger, A. (2008). Multicultural assessment of child and adolescent psychopatholgy with ASEBA and SDQ instruments: research findings, applications, and future directions. *Journal of Child Psychology and Psychiatry 49*, 251-275.

Achenbach, T. M., & Dumenci, L. (2001). Advances in empirically based assessment: Revised cross-informant syndromes and new DSM-oriented scales for the CBCL, YSR, and TRF: Comment on Lengua, Sadowski, Friedrich, and Fisher (2001). *Journal Of Consulting And Clinical Psychology*, *69*(4), 699-702. doi:10.1037/0022-006X.69.4.699

Achenbach, T. M., Dumenci, L., & Rescorla, L. A. (2003). Are American children's problems still getting worse? A 23-year comparison. *Journal of Abnormal Child Psychology, 31*, 1-11. doi:10.1023/A:1021700430364

Achenbach, T. M., & Edelbrock, C. (1983). *Manual for the Child Behavior Checklist and Revised Child Behavior Profile*. Burlington: University of Vermont, Department of Psychiatry.

Achenbach, T. M., Krukowski, R. A., Dumenci, L., & Ivanova, M. Y. (2005). Assessment of adult psychopathology: Meta-analyses and implications of cross-informant correlations. *Psychological Bulletin, 131*, 361-382. doi:2005-04167-003 [pii]10.1037/0033-2909.131.3.361

Achenbach, T. M., McConaughy, S. H., & Howell, C. T. (1987). Child/Adolescent behavioral and emotional problems: Implication of cross-informant correlations for situational specificity. *Psychological Bulletin, 101*, 213-232. doi: 10.1037/0033-2909.101.2.213

Achenbach, T. M., & Rescorla, L. A. (2001). *Manual for the ASEBA School-Age Forms & Profiles*. Burlington, VT: University of Vermont, Research Center for Children, Youth, & Families.

Achenbach, T. M., & Rescorla, L. A. (2007a). *Multicultural Supplement to the Manual for the ASEBA School-Age Forms & Profiles*. Burlington, VT: University of Vermont, Research Center for Children, Youth, & Families.

Achenbach, T. M., & Rescorla, L. A. (2007b). *Multicultural Understanding of Child and Adolescent Psychopathology: Implications for Mental Health Assessment*. New York, NY: Guilford Press.

Achenbach, T.M., Rescorla, L.A., & Ivanova, M.Y. (2012). International epidemiology of child and adolescent psycho-pathology: 1. Diagnoses, dimensions, and conceptual issues. *Journal of the American Academy of Child and Adolescent Psychiatry, 51*, 1261-1272.

Albores-Gallo, L., Lara-Muñoz, C., Esperón-Vargas, C., Zetina, J. C., Soriano, A. P., & Colin, G. V. (2007). Validity and reliability of the CBCL/6-18. Includes DSM scales. *Actas Españolas De Psiquiatría, 35*(6), 393-399.

Althoff, R. R., Ayer, L. A., Rettew, D. C., & Hudziak, J. J. (2010). Assessment of dysregulated children using the Child Behavior Checklist: a receiver operating characteristic curve analysis. *Psychological Assessment, 22*, 609-617. doi: 10.1037/a0019699

American Psychiatric Association. (1987). *Diagnostic and statistical manual of mental disorders* (3rd ed., revised ed.). Washington, DC: American Psychiatric Association.

American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.). Washington, DC: American Psychiatric Association.

American Psychiatric Association. (2000). *Diagnostic and statistical manual of mental disorders* (4th ed., revised ed.). Washington, DC: American Psychiatric Association.

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). Arlington, VA: American Psychiatric Association.

APA Publications and Communications Board Working Group on Journal Article Reporting Standards. (2008). Reporting standards for research in psychology: why do we need them? What might they be? *American Psychologist, 63*, 839-851. doi:10.1037/0003-066X.63.9.839

Aschenbrand, S. G., Angelosante, A. G., & Kendall, P. C. (2005). Discriminant validity and clinical utility of the CBCL with anxiety-disordered youth. *Journal of Clinical Child and Adolescent Psychology, 34*(4), 735-746. doi:10.1207/s15374424jccp3404_15

Beidas, R. S., Stewart, R. E., Walsh, L., Lucas, S., Downey, M. M., Jackson, K., . . . Mandell, D. S. (2015). Free, brief, and validated: Standardized instruments for low-resource mental health settings. *Cognitive & Behavioral Practice, 22*, 5-19. doi:10.1016/j.cbpra.2014.02.002

Bellina, M., Brambilla, P., Garzitto, M., Negri, G. L., Molteni, M., & Nobile, M. (2013). The ability of CBCL DSM-oriented scales to predict DSM-IV diagnoses in a referred sample of children and adolescents. *European Child & Adolescent Psychiatry*, *22*(4), 235-246. doi:10.1007/s00787-012-0343-0

Berkey, C.S., Hoaglin, D.C., Antczak-Bouckoms, A., Mosteller, F., & Colditz, G. A. (1998). Meta-analysis of multiple outcomes by regression with random effects. *Stat Med, 17*(22), 2537-2550

Biederman, J., Monuteauz, M. C., Kendrick, E., Klein, K. L., & Faraone, S. V. (2005). The CBCL as a screen for psychiatric comorbidity in paedriatric patients with ADHD. *Archives of Disease Childhood, 90,* 1010-105. doi: 10.1136/adc.2004.056937

Biederman, J., Wozniak, J., Kiely, K., Ablon, S., Faraone, S., Mick, E., . . . Kraus, I. (1995). CBCL clinical scales discriminate prepubertal children with structured interview-derived diagnosis of mania from those with ADHD. *Journal of the American Academy of Child & Adolescent Psychiatry, 34*, 464-471.

Bird, H. R., Gould, M. S., Rubio-Stipec, M., Staghezza, B. M., & Canino, G. (1991). Screening for childhood psychopathology in the community using the Child Behavior Checklist. *Journal of the American Academy of Child & Adolescent Psychiatry, 30*(1), 116–123. https://doi.org/10.1097/00004583-199101000-00018

Bossuyt, P. M., Reitsma, J. B., Bruns, D. E., Gatsonis, C. A., Glasziou, P. P., Irwig, L. M., . . . de Vet, H. C. W. (2003). Towards complete and accurate reporting of studies of diagnostic accuracy: The STARD initiative. *British Medical Journal, 326*, 41-44. doi:10.1136/bmj.326.7379.41

Bossuyt, P. M., Reitsma, J. B., Bruns, D. E., Gatsonis, C. A., Glasziou, P. P., Irwig, L. M., . . . Lijmer, J. G. (2003). The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration. *Clinical Chemistry, 49*, 7-18.

Camara, W. J., Nathan, J. S., & Puente, A. E. (2000). Psychological test usage: Implications in professional psychology. *Professional Psychology: Research and Practice*, *31*, 141.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by multitrait-multimethod matrix. *Psychological Bulletin*, *56*, 81-105.

Canning, E. H., & Kelleher, K. (1994). Performance of sccreening tools for mental health problems in chronically ill children. *Arch Pediatr Adolesc Med, 148*, 272-278.

Carlson, G. A., & Youngstrom, E. A. (2011). Two opinions about one child-What's the clinician to do? *Journal of Child and Adolescent Psychopharmacology, 21*, 385-387. doi: 10.1089/cap.2011.2159

Choudhry, F. R., Mani, V., Ming, L. C., & Khan, T. M. (2016). Beliefs and perception about mental health issues: a meta-synthesis. *Neuropsychiatric disease and treatment*, *12*, 2807–2818. https://doi.org/10.2147/NDT.S111543

Clarizio, H. F. (1994). Assessment of depression in children and adolescents by parents, teachers, and peers. In W. M. Reynolds, H. F. Johnston, W. M. Reynolds, H. F. Johnston (Eds.) , *Handbook of depression in children and adolescents*(pp. 235-248). New York, NY, US: Plenum Press. doi:10.1007/978-1-4899-1510-8_12

Clarke, G. N., Lewinsohn, P. M., Hops, H., & Seeley, J. R. (1992). A self- and parent-report measure of adolescent depression: The Child Behavior Checklist Depression scale (CBCL-D). *Behavioral Assessment*, *14*(3-4), 443-463.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Curry, J. F., & Ilardi, S. S. (2000). Validity of the devereux scales of mental disorders withadolescent psychiatric inpatients. *Journal of Clinical Child Psychology, 29*(4), 578-588. doi:10.1207/S15374424JCCP2904_10

De Los Reyes, A., & Kazdin, A. E. (2005). Informant discrepancies in the assessment of childhood psychopathology: A critical review, theoretical framework, and recommendations for further study. *Psychological Bulletin, 131*, 483-509. doi: 10.1037/0033-2909.131.4.483

DelBello, M. P., Lopez-Larson, M. P., Soutullo, C. A., & Strakowski, S. M. (2001). Effects of race on psychiatric diagnosis of hospitalized adolescents: A retrospective chart review. *Journal of Child and Adolescent Psychopharmacology*, *11*, 95-103.

Dell'Osso, L., Pini, S., Cassano, G. B., Mastrocinque, C., Seckinger, R. A., Saettoni, M., . . . Amador, X. F. (2002). Insight into illness in patients with mania, mixed mania, bipolar depression and major depression with psychotic features. *Bipolar Disorders, 4*, 315-322.

Derks, E. M., Hudziak, J. J., Dolan, C. V., Ferdinand, R. F., & Boomsma, D. I. (2006). The relations between DISC-IV DSM diagnoses of ADHD and multi-informant CBCL-AP syndrome scores. *Comprehensive Psychiatry*, *47*(2), 116-122. doi:10.1016/j.comppsych.2005.05.006

Dienes, K. A., Chang, K. D., Blasey, C. M., Adleman, N., & Steiner, H. (2002). Characterization of children of bipolar parents by parent report CBCL. *Journal of Psychiatric Research*, *36*, 337-345.

Diler, R. S., Birmaher, B., Axelson, D., Goldstein, B., Gill, M., Strober, M., . . . Keller, M. B. (2009). The Child Behavior Checklist (CBCL) and the CBCL-bipolar phenotype are not useful in diagnosing pediatric bipolar disorder. *Journal of Child and Adolescent Psychopharmacology, 19*, 23-30. doi: 10.1089/cap.2008.067

Doerfler, L. A., Connor, D. F., & Toscano, P. F. (2010). The CBCL bipolar profile and attention, mood, and behavior dysregulation. *Journal of Child and Family Studies*, *20*, 545-553. doi: 10.1007/s10826-010-9426-z

Doerfler, L. A., Connor, D. F., & Toscano, P. F., Jr. (2011). Aggression, ADHD symptoms, and dysphoria in children and adolescents diagnosed with bipolar disorder and ADHD. *Journal of Affective Disorders, 131*, 312-319. doi: 10.1016/j.jad.2010.11.029

Doyle, A., Ostrander, R., Skare, S., Crosby, R. D., & August, G. J. (1997). Convergent and criterion-related validity of the behavior assessment system for children-parent rating scale. *Journal of Clinical Child Psychology, 26(*3), 276-284. doi: 10.1207/s15374424jccp2603_6

Drotar, D., Stein, R. E. K., & Perrin, E. C. (1995). Methodological issues in using the Child Behavior Checklist and its related instruments in clinical child psychology research. *Journal of Clinical Child Psychology, 24*, 184-192.

Edwards, M. C., & Sigel, B. A. (2015). Estimates of the utility of child behavior checklist/teacherreport form attention problems scale in the diagnosis of adhdin children referred to a specialty clinic. *J Psychopathol Behav Assess, 37*, 50–59. doi: 10.1007/s10862-014-9431-4

Eimecke, S. D., Remschmidt, H., & Mattejat, F. (2011). Utility of the child behavior checklist in screening depressive disorderswithin clinical samples. *Journal of Affective Disorders, 129*, 191–197.

Eiraldi, R. B., Power, T. J., Karustis, J. L., & Goldstein, S. G. (2000). Assessing adhdand comorbid disorders in children: the child behavior checklist and the devereux scales of mental disorders. *Journal of Clinical Child Psychology, 29*(1), 3-16. doi: 10.1207/S15374424jccp2901_2

Elkins, R. M., Carpenter, A. L., Pincus, D. B., & Comer, J. S. (2014). Inattention symptoms and the diagnosis of comorbid attention-deficit/hyperactivity disorder among youth with generalized anxiety disorder. *Journal of Anxiety Disorders, 28*, 754-760.

Embretson, S. E. (1996). The new rules of measurement. *Psychological Assessment, 8,* 341-349.

Faraone, S. V., Althoff, R. R., Hudziak, J. J., Monuteaux, M., & Biederman, J. (2005). The CBCL predicts DSM bipolar disorder in children: A receiver operating characteristic curve analysis. *Bipolar Disorders, 7*, 518-524.

Freeman, A. J., Youngstrom, E. A., Freeman, M. J., Youngstrom, J. K., & Findling, R. L. (2011). Is caregiver-adolescent disagreement due to differences in thresholds for reporting manic symptoms? *Journal of Child and Adolescent Psychopharmacology, 21*, 425-432. doi:10.1089/cap.2011.0033

Geller, B., Warner, K., Williams, M., & Zimerman, B. (1998). Prepubertal and young adolescent bipolarity versus ADHD: assessment and validity using the WASH-U-KSADS, CBCL and TRF. *Journal of Affective Disorders, 51*, 93-100.

Gleser, L. J., & Olkin, I. (2009). Stochastically dependent effect sizes. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed., pp. 357-376). New York, NY: Sage.

Gopalkrishnan, N., & Babacan, H. (2015). Cultural diversity and mental health. *Australasian psychiatry : bulletin of Royal Australian and New Zealand College of Psychiatrists*, *23*(6 Suppl), 6–8. https://doi-org.libproxy.lib.unc.edu/10.1177/1039856215609769

Gore, F. M., Bloem, P. J., Patton, G. C., Ferguson, J., Joseph, V., Coffey, C., . . . Mathers, C. D. (2011). Global burden of disease in young people aged 10-24 years: a systematic analysis. *Lancet, 377,* 2093-2102. doi: 10.1016/S0140-6736(11)60512-6

Hasselbad, V., & Hedges, L. V. (1995). Meta-analysis of screening and diagnostic tests. *Psychological Bulletin, 117*, 167-178. doi:10.1037/0033-2909.117.1.167

Havdahl, K. A., von Tetzchner, S., Huerta, M., Lord, C., & Bishop, S. L. (2016). Utility of the child behavior checklist as a screener for autism spectrum disorder. *Autism Research*, *9*(1), 33-42. doi:10.1002/aur.1515

Hayatbakhsh, M. R., Najman, J. M., Jamrozik, K., Mamun, A. A., Bor, W., & Alati, R. (2008). Adolescent problem behaviours prediting dsm-iv diagnoses of multiple substance use disorder. *Psychiatr Epidemiol, 43,* 356-363.

Hazell, P. L., Lewin, T. J., & Carr, V. J. (1999). Confirmation that Child Behavior Checklist clinical scales discriminate juvenile mania from attention deficit hyperactivity disorder. *Journal of Paediatrics and Child Health, 35*, 199-203.

Hedges, L.V., & Olkin, I. (1985). *Statistical Methods for Meta-Analysis.* San Diego, CA: Academic Press

Hedges, L. V., & Pigott, T. D. (2001). The power of statistical tests in meta-analysis. *Psychological Methods, 6,* 203-217.

Hofstra, M. B., van der Ende, J., & Verhulst, F. C. (2002). Child and adolescent problems predict DSM-IV disorders in adulthood: A 14-year follow-up of a Dutch epidemiological sample. *Journal of the American Academy of Child and Adolescent Psychiatry, 41,* 182-189.

Hummel, T. J. (1999). The usefulness of tests in clinical decisions. In J. W. Lichtenberg & R. K. Goodyear (Eds.), *Scientist-practitioner perspectives on test interpretation* (pp. 59-112). Boston, MA: Allyn and Bacon.

Ivanova, M.Y., Achenbach, T.M., Dumenci, L., Rescorla, L.A., Almqvist, F., Bilenberg, N., et al. (2007a). Testing the 8-syndrome structure of the Child Behavior Checklist in 30 societies. *Journal of Clinical Child and Adolescent Psychology, 36,* 405-417.

Ivanova, M.Y., Achenbach, T.M., Rescorla, L.A., Dumenci, L., Almqvist, F., Bathiche, M. et al. (2007b). Testing the Teacher's Report Form syndromes in 20 societies. *School Psychology Review, 36,* 468-483.

Ivanova, M.Y., Achenbach, T.M., Rescorla, L.A., Dumenci, L., Almqvist, F., Bilenberg, N., et al. (2007c). The generalizability of the Youth Self-Report syndrome structure in 23 societies. *Journal of Consulting and Clinical Psychology, 75,* 729-738.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning: with applications in R*. New York, NY: Springer.

Jenkins, M. M., Youngstrom, E. A., Washburn, J. J., & Youngstrom, J. K. (2011). Evidence-based strategies improve assessment of pediatric bipolar disorder by community practitioners. *Professional Psychology: Research and Practice, 42*, 121-129. doi: 10.1037/a0022506

Jenkins, M. M., Youngstrom, E. A., Youngstrom, J. K., Feeny, N. C., & Findling, R. L. (2012). Generalizability of evidence-based assessment recommendations for pediatric bipolar disorder. *Psychological Assessment, 24*, 269-281. doi: 10.1037/a0025775

Jensen-Doss, A., Youngstrom, E. A., Youngstrom, J. K., Feeny, N. C., & Findling, R. L. (2014). Predictors and moderators of agreement between clinical and research diagnoses for children and adolescents. *Journal of Consulting & Clinical Psychology, 82*, 1151-1162. doi:10.1037/a0036657

Kahana, S. Y., Youngstrom, E. A., Findling, R. L., & Calabrese, J. R. (2003). Employing parent, teacher, and youth self-report checklists in identifying pediatric bipolar spectrum disorders: an examination of diagnostic accuracy and clinical utility. *Journal of Child and Adolescent Psychopharmacology, 13*, 471-488.

Kasius, M. C., Ferdinand, R. F., van den Berg, H., & Verhulst, F. C. (1997). Associations between different diagnostic approaches for child and adolescent psychopathology. *Child Psychology & Psychiatry & Allied Disciplines*, *38*(6), 625-632. doi:10.1111/j.1469-7610.1997.tb01689.x

Kim, J., Park, K, Cheon, K., Kim, B., Cho, S., & Hong, K. M. (2015). The child behavior checklist together with the adhd rating scale can diagnose adhd in korean community-based samples. *Canadian Journal of Psychiatry, 50*(12), 802.

Konstantopoulous, S. (2011). Fixed effects and variance components estimation in three-level meta-analysis. *Research Synthesis Methods, 2*, 61-76. doi:10.1002/jrsm.35

Kowatch, R. A., Youngstrom, E. A., Danielyan, A., & Findling, R. L. (2005). Review and meta-analysis of the phenomenology and clinical characteristics of mania in children and adolescents. *Bipolar Disorders, 7*, 483-496. doi:10.1111/j.1399-5618.2005.00261.x

Kraemer, H. C. (1992). *Evaluating medical tests: Objective and quantitative guidelines*. Newbury Park, CA: Sage.

Krol, N. M., De Bruyn, E. J., Coolen, J. C., & van Aarle, E. M. (2006). From CBCL to DSM: A Comparison of Two Methods to Screen for DSM-IV Diagnoses Using CBCL Data. *Journal Of Clinical Child And Adolescent Psychology, 35*(1), 127-135. doi:10.1207/s15374424jccp3501_11

Lengua, L. J., Sadowski, C. A., Friedrich, W. N., & Fisher, J. (2001). Rationally and empirically derived dimensions of children's symptomatology: Expert ratings and confirmatory factor analyses of the CBCL. *Journal Of Consulting And Clinical Psychology, 69*(4), 683-698. doi:10.1037/0022-006X.69.4.683

Liberati, A., Altman, D. G., Tetzlaff, J., Mulrow, C., Gotzsche, P. C., Ioannidis, J. P., . . . Moher, D. (2009). The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions: Explanation and elaboration. *BMJ, 339*, b2700. doi:10.1136/bmj.b2700

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis* (Vol. 49). Thousand Oaks, CA: Sage Publications.

Loeber, R., Green, S. M., & Lahey, B. B. (1990). Mental health professionals' perception of the utility of children, mothers, and teachers as informants on childhood psychopathology. *Journal of Clinical Child Psychology, 19*, 136-143.

MacDonald, V. M., & Achenbach, T. M. (1996). Attention problems versus conduct problems as six-year predictors of problem scores in a national sample. *Journal of the American Academy of Child and Adolescent Psychiatry, 35,* 1237-1246.

MacDonald, V.M., & Achenbach, T.M. (1999). Attention problems versus conduct problems as 6-year predictors of signs of disturbance in a national sample. *Journal of the American Academy of Child and Adolescent Psychiatry, 38,* 1254-1261.

Mash, E. J., & Barkley, R. A. (2014). *Child psychopathology., 3rd ed*. New York, NY, US: Guilford Press.

Mbekou, V., Gignac, M., MacNeil, S., Mackay, P., & Renaud, J. (2014). The CBCL dysregulated profile: an indicator of pediatric bipolar disorder or of psychopathology severity? *Journal of Affective Disorders, 155,* 299-302. doi: 10.1016/j.jad.2013.10.033

Meeuwesen, L., van den Brink-Muinen, A., & Hofstede, G. (2009). Can dimensions of national culture predict cross-national differences in medical communication? *Patient education and counseling, 75*, 58-66. doi:10.1016/j.pec.2008.09.015

Merikangas, K. R., He, J. P., Burstein, M., Swendsen, J., Avenevoli, S., Case, B., . . . Olfson, M. (2011). Service utilization for lifetime mental disorders in U.S. adolescents: Results of the National Comorbidity Survey-Adolescent Supplement (NCS-A). *Journal of the American Academy of Child and Adolescent Psychiatry, 50,* 32-45. doi: 10.1016/j.jaac.2010.10.006

Meyer, G. J. (2003). Guidelines for reporting information in studies of diagnostic test accuracy: The STARD initiative. *Journal of Personality Assessment, 81*, 191-193.

Meyer, S. E., Carlson, G. A., Youngstrom, E., Ronsaville, D. S., Martinez, P. E., Gold, P. W., . . . Radke-Yarrow, M. (2009). Long-term outcomes of youth who manifested the CBCL-Pediatric Bipolar Disorder phenotype during childhood and/or adolescence. *Journal of Affective Disorders, 113*, 227-235. doi: S0165-0327(08)00244-9 [pii] 10.1016/j.jad.2008.05.024

Mick, E., Biederman, J., Pandina, G., & Faraone, S. V. (2003). A preliminary meta-analysis of the child behavior checklist in pediatric bipolar disorder. *Biological Psychiatry, 53*, 1021-1027. doi: 10.1016/S0006-3223(03)00234-8 |

Miller, C. J., Johnson, S. L., Kwapil, T. R., & Carver, C. S. (2011). Three studies on self-report scales to detect bipolar disorder. *Journal of Affective Disorders, 128*, 199-210. doi: 10.1016/j.jad.2010.07.012

Morrison, J. (2007). *Diagnosis Made Easier: Principles and Techniques for Mental Health Clinicians*. New York, NY: Guilford Press.

Najman, J. M., Heron, M. A., Hayatbakhsh, M. R., Dingle, K., Jamrozik, K., Bor, W., O'Callaghan, M. J., & Williams, G. M. (2008). Screening in early childhood for risk of later mental health problems: a longitudinal study. *Journal of Psychiatric Research, 42,* 694-709.

Nobile, M., Colombo, P., Bellina, M., Molteni, M., Simone, D., Nardocci, F., & ... Battaglia, M. (2013). Psychopathology and adversities from early- to late-adolescence: A general population follow-up study with the CBCL DSM-Oriented Scales. *Epidemiology And Psychiatric Sciences, 22*(1), 63-73. doi:10.1017/S2045796012000145

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*, aac4716-aac4716. doi:10.1126/science.aac4716

Orwin, R. G. (1983). A fail-safe *N* for effect size in meta-analysis. *Journal of Educational Statistics, 8,* 157-159.

Ostrander, R., Weinfurt, K. P., Yarndold, P. R., & August, G. J. (1998). Diagnosing attention deficit disorders with the behavioral assessment system for children and the child behavior checklist: test and construct validity analyses using optimal discriminant classification trees. *Journal of Counseling and Clinical Psychology*, *66*(4), 660-672.

Papachristou, E., Ormel, J., Oldehinkel, A. J., Kyriakopoulos, M., Reinares, M., Reichenberg, A., & Frangou, S. (2013). Child Behavior Checklist-Mania Scale (CBCL-MS): development and evaluation of a population-based screening scale for bipolar disorder. *PLoS One*, *8*, e69459. doi: 10.1371/journal.pone.0069459

Park, J., Shim, S. Lee, M., Jung, Y., Park, T. W., Park, S. H., Im, Y., Yang, J., Chung, Y., & Chung, S. (2014). The validities and efficiencies of korean adhd ratingscale and korean child behavior checklist for screeningchildren with adhd in the community. *Psychiatry Investig*, *11*(3), 256-265.

Pepe, M. S. (2003). *The statistical evaluation of medical tests for classification and prediction*. New York, NY: Wiley.

Pesenti-Gritti, P., Scaini, S., D'ippolito, C., Fagnani, C., & Battaglia, M. (2011). A genetically informed study of the covariation between the CBCL/6-18 DSM-Oriented Problem Scales and the Competence Scales. *Behavior Genetics*, *41*(4), 522-532. doi:10.1007/s10519-010-9420-7

Podsakoff, P. M., MacKenzie, S. B., & Podsakoff, N. P. (2012). Sources of method bias in social science research and recommendations on how to control it. *Annual Review of Psychology*, *63,* 539-569. doi: 10.1146/annurev-psych-120710-100452

R Core Team (2014). *R: A language and environment for statistical computing.* Vienna, Austria.

Reef, J., Diamantopoulou, S., van Meurs, I., Verhulst, F., & van der Ende, J. (2010a). Predicting adult emotional and behavioral problems from externalizing problem trajectories in a 24-year longitudinal study. *Journal of European Child and Adolescent Psychiatry, 19,* 577-585.

Reef, J., van Meurs, I., Verhulst, F.C., & van der Ende, J. (2010b). Children's problems predict adults' DSM-IV disorders across 24 years. *Journal of the American Academy of Child and Adolescent Psychiatry, 49,* 1117-1124.

Rescorla, L.A., Ivanova, M.Y., Achenbach, T.M., Begovac, I., Chahed, M., Drugli, M.B., …Zhang, E.Y. (2012). International epidemiology of child and adolescent psychopathology: 2. Integration and applications of dimensional findings from 44 societies. *Journal of the American Academy of Child and Adolescent Psychiatry, 51*, 1273-1283.

Rettew, D. C., Lynch, A. D., Achenbach, T. M., Dumenci, L., & Ivanova, M. Y. (2009). Meta-analyses of agreement between diagnoses made from clinical evaluations and standardized diagnostic interviews. *International Journal of Methods in Psychiatric Research, 18,* 169-184. doi:10.1002/mpr.289

Rey, J. M., & Morris-Yates, A. (1991). Adolescent depression and the Child Behavior Checklist. *Journal Of The American Academy Of Child & Adolescent Psychiatry*, *30*(3), 423-427. doi:10.1097/00004583-199105000-00011

Rice, M. E., & Harris, G. T. (2005). Comparing effect sizes in follow-up studies: ROC Area, Cohen's *d*, and *r*. *Law and Human Behavior*, *29*, 615-620.

Richters, J. E. (1992). Depressed mothers as informants about their children: A critical review of the evidence for distortion. *Psychological Bulletin*, *112*, 485-499. doi: 10.1037//0033-2909.112.3.485

Roessner, V., Becker, A., Rothenberger, A., Rohde, L. A., & Banaschewski, T. (2007). A cross-cultural comparison between samples of Brazilian and German children with ADHD/HD using the Child Behavior Checklist. *Clin Neurosci*, *257*, 352-359.

Rosenberg, M. S. (2005). The file-drawer problem revisited: a general weighted method for calculating fail-safe numbers in meta-analysis. *Evolution*, *59*, 464-468.

Rucklidge, J. J. (2008). Retrospective parent report of psychiatric histories: Do checklists reveal specific prodromal indicators for postpubertal-onset pediatric bipolar disorder? *Bipolar Disorders*, *10*, 56-66.

Sattler, J. M. (2002). *Assessment of children: Behavioral and Clinical Applications* (4th ed.). La Mesa, CA: Author.

Schwarz, N. (1999). Self-reports: How the questions shape the answers. *American Psychologist*, *54*, 93-105.

Šimundić, A.-M. (2009). Measures of diagnostic accuracy: Basic definitions. *EJIFCC*, *19*(4), 203–211.

Song, L., Singh, J., & Singer, M. (1994). The Youth Self-Report inventory: A study of its measurements fidelity. *Psychological Assessment*, *6*(3), 236-245. doi:10.1037/1040-3590.6.3.236

Southammakosane, C., Danielyan, A., Welge, J. A., Blom, T. J., Adler, C. M., Chang, K. D., & ... DelBello, M. P. (2013). Characteristics of the Child Behavior Checklist in adolescents with depression associated with bipolar disorder. *Journal Of Affective Disorders*, *145*(3), 405-408. doi:10.1016/j.jad.2012.06.017

Spatola, C. M., Fagnani, C., Pesenti-Gritti, P., Ogliari, A., Stazi, M., & Battaglia, M. (2007). A general population twin study of the CBCL/6-18 DSM-oriented scales. *Journal Of The American Academy Of Child & Adolescent Psychiatry*, *46*(5), 619-627. doi:10.1097/CHI.0b013e3180335b12

Spatola, C. M., Rende, R., & Battaglia, M. (2010). Genetic and environmental influences upon the CBCL/6-18 DSM-oriented scales: Similarities and differences across three different computational approaches and two age ranges. *European Child & Adolescent Psychiatry*, *19*(8), 647-658. doi:10.1007/s00787-010-0102-z

Spitzer, R. L. (1983). Psychiatric diagnosis: Are clinicians still necessary? *Comprehensive Psychiatry, 24*, 399-411.

Stedman, J. M., Hatch, J. P., & Schoenfeld, L. S. (2001). The current status of psychological assessment training in graduate and professional schools. *Journal of Personality Assessment, 77*, 398-407. doi: 10.1207/S15327752JPA7703_02

Straus, S. E., Glasziou, P., Richardson, W. S., & Haynes, R. B. (2011). *Evidence-based medicine: How to practice and teach EBM* (4th ed.). New York, NY: Churchill Livingstone.

Swets, J. A., Dawes, R. M., & Monahan, J. (2000). Psychological science can improve diagnostic decisions. *Psychological Science in the Public Interest, 1,* 1-26. doi: 10.1111/1529-1006.001

Tripp, G., Schaughency, E. A., & Clarke, B. (2006). Parent and teacher rating scales in the evaluationof attention-deficit hyperactivity disorder:contribution to diagnosis and differential diagnosisin clinically referred children. *Developmental and Behavioral Pediatrics, 27*(3). doi: 0196-206X/06/2703-0209

Uchida, M., Faraone, S. V., Martelon, M., Kenworthy, T., Woodworth, K. Y., Spencer, T. J., . . . Biederman, J. (2014). Further evidence that severe scores in the aggression/anxiety-depression/attention subscales of child behavior checklist (severe dysregulation profile) can screen for bipolar disorder symptomatology: a conditional probability analysis. *Journal of Affective Disorders, 165*, 81-86. doi: 10.1016/j.jad.2014.04.021

Van Meter, A., Youngstrom, E., Youngstrom, J. K., Ollendick, T., Demeter, C., & Findling, R. L. (2014). Clinical decision making about child and adolescent anxiety disorders using the Achenbach System of Empirically Based Assessment. *Journal of Clinical Child & Adolescent Psychology, 43*, 552-565. doi: 10.1080/15374416.2014.883930

Van Oort, F.V.A., van der Ende, J., Wadsworth, M.E., Verhulst, F.C., & Achenbach, T.M. (2011). Cross-national comparison of the link between socioeconomic status and emotional and behavioral problems in youths. *Social Psychiatry and Psychiatric Epidemiology, 46*, 167-172. doi:10.1007/s00127-010-0191-5

Venkatraman, E. S. (2000). A permutation test to compare receiver operating characteristic curves. *Biometrics*, *56*, 1134-1138.

Viechtbauer, W. (2007). Hypothesis tests for population heterogeneity in meta-analysis. *The British Journal of Mathematical and Statistical Psychology, 60*, 29-60. doi:

10.1348/000711005X64042

Viechtbauer, W. (2010a). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software, 36*, 1-48.

Viechtbauer, W. (2010b). Metafor: meta-analysis package for R. *R package version, 2010*, 1-0.

Wadsworth, M. E., Hudziak, J. J., Heath, A. C., & Achenbach, T. M. (2001). Latent class analysis of Child Behavior Checklist anxiety/depression in children and adolescents. *Journal Of The American Academy Of Child & Adolescent Psychiatry*, *40*(1), 106-114. doi:10.1097/00004583-200101000-00023

Waugh, M. J., Meyer, T. D., Youngstrom, E. A., & Scott, J. (2014). A review of self-rating instruments to identify young people at risk of bipolar spectrum disorders. *Journal of Affective Disorders, 160*, 113-121. doi: 10.1016/j.jad.2013.12.019

Whiting, P., Rutjes, A. W., Reitsma, J. B., Bossuyt, P. M., & Kleijnen, J. (2003). The development of QUADAS: A tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. *BMC Medical Research Methodology, 3*, 25.

Whiting, P. F., Rutjes, A. W., Westwood, M. E., Mallett, S., Deeks, J. J., Reitsma, J. B., . . . Bossuyt, P. M. (2011). QUADAS-2: A revised tool for the quality assessment of diagnostic accuracy studies. *Annals of Internal Medicine, 155*, 529-536. doi:10.7326/0003-4819-155-8-201110180-00009

World Health Organization. (1992). *The ICD-10 Classification of Mental and Behavioural Disorders: Clinical Descriptions and Diagnostic Guidelines*. London, UK: World Health Organization.

World Health Organization. (2009). *Global health risks: Mortality and burden of disease attributable to selected major risks*. Geneva, Switzerland: Author.

Yeh, M., & Weisz, J. (2001). Why are we here at the clinic? Parent-child (dis)agreement on referral problems at outpatient treatment entry. *Journal of Consulting and Clinical Psychology*, *69*, 1018-1025.

You, D. S., Youngstrom, E. A., Feeny, N. C., Youngstrom, J. K., & Findling, R. L. (2015). Comparing the diagnostic accuracy of five instrumentsfor detecting posttraumatic stress disorder in youth. *Journal of Clinical Child & Adolescent Psychology*, 1-12. doi: 10.1080/15374416.2015.1030754

Youngstrom, E. A. (2009). Definitional issues in bipolar disorder across the life cycle. *Clinical Psychology: Science & Practice, 16*, 140-160. doi:10.1111/j.1468-2850.2009.01154.x

Youngstrom, E. A. (2013). Future directions in psychological assessment: Combining Evidence-Based Medicine innovations with psychology's historical strengths to enhance utility.

Journal of Clinical Child & Adolescent Psychology, 42, 139-159. doi:10.1080/15374416.2012.736358

Youngstrom, E. A. (2014). A primer on Receiver Operating Characteristic analysis and diagnostic efficiency statistics for pediatric psychology: We are ready to ROC. *Journal of Pediatric Psychology, 39*, 204-221. doi: 10.1093/jpepsy/jst062

Youngstrom, E. A., Choukas-Bradley, S., Calhoun, C. D., & Jensen-Doss, A. (2014). Clinical guide to the Evidence-Based Assessment approach to diagnosis and treatment. *Cognitive and Behavioral Practice, 22*, 20-35. doi: 10.1016/j.cbpra.2013.12.005

Youngstrom, E. A., Egerton, G., Genzlinger, J., Rizvi, S. H., Freeman, L. K., & Van Meter, A. (2018). Improving the global identification of bipolar spectrum disorders: Meta-analysis of the diagnostic accuracy of checklists. *Psychological Bulletin*.

Youngstrom, E. A., Findling, R. L., & Calabrese, J. R. (2003). Who are the comorbid adolescents? Agreement between psychiatric diagnosis, parent, teacher, and youth report. *Journal of Abnormal Child Psychology, 31*, 231-245.

Youngstrom, E. A., Findling, R. L., Calabrese, J. R., Gracious, B. L., Demeter, C., DelPorto Bedoya, D., & Price, M. (2004). Comparing the diagnostic accuracy of six potential screening instruments for bipolar disorder in youths aged 5 to 17 years. *Journal of the American Academy of Child & Adolescent Psychiatry, 43*, 847-858. doi: 10.1097/01.chi.0000125091.35109.1e

Youngstrom, E. A., & Frazier, T. W. (2013). Evidence-based strategies for the assessment of children and adolescents: Measuring prediction, prescription, and process. In D. J. Miklowitz, W. E. Craighead, & L. Craighead (Eds.), *Developmental psychopathology* (2nd ed., pp. 36-79). New York, NY: Wiley.

Youngstrom, E. A., Frazier, T. W., Findling, R. L., & Calabrese, J. R. (2008). Developing a ten item short form of the Parent General Behavior Inventory to assess for juvenile mania and hypomania. *Journal of Clinical Psychiatry, 69*, 831-839. doi:10.4088/JCP.v69n0517

Youngstrom, E. A., Genzlinger, J. E., Egerton, G. A., & Van Meter, A. R. (2015). Multivariate meta-analysis of the discriminative validity of caregiver, youth, and teacher rating scales for pediatric bipolar disorder: Mother knows best about mania. *Archives of Scientific Psychology, 3*, 112-137. doi:10.1037/arc0000024

Youngstrom, E. A., Joseph, M. F., & Greene, J. (2008). Comparing the psychometric properties of multiple teacher report instruments as predictors of bipolar disorder in children and adolescents. *Journal of Clinical Psychology, 64*, 382-401.

Youngstrom, E. A., Meyers, O. I., Demeter, C., Kogos Youngstrom, J., Morello, L., Piiparinen, R., . . . Calabrese, J. R. (2005). Comparing diagnostic checklists for pediatric bipolar disorder in academic and community mental health settings. *Bipolar Disorders, 7,* 507-517. doi:10.1111/j.1399-5618.2005.00269.x

Youngstrom, E. A., Meyers, O. I., Youngstrom, J. K., Calabrese, J. R., & Findling, R. L. (2006). Comparing the effects of sampling designs on the diagnostic accuracy of eight promising screening algorithms for pediatric bipolar disorder. *Biological Psychiatry, 60,* 1013-1019. doi:10.1016/j.biopsych.2006.06.023

Youngstrom, E. A., Murray, G., Johnson, S. L., & Findling, R. L. (2013). The 7 Up 7 Down Inventory: A 14-item measure of manic and depressive tendencies carved from the General Behavior Inventory. *Psychological Assessment, 25*, 1377-1383. doi:10.1037/a0033975

Youngstrom, E. A., & Van Meter, A. (2013). Comorbidity of bipolar disorder and depression. In S. Richards & M. W. O'Hara (Eds.), *Oxford Handbook of Depression and Comorbidity*. New York, NY: Oxford University Press.

Youngstrom, E. A., Van Meter, A., Frazier, T. W., Hunsley, J., Prinstein, M., Ong, M.-L., & Youngstrom, J. K. (in press). Evidence-based assessment as an integrative model for applying psychological science to guide the voyage of treatment. *Clinical Psychology: Science & Practice*.

Youngstrom, E. A., Youngstrom, J. K., Freeman, A. J., De Los Reyes, A., Feeny, N. C., & Findling, R. L. (2011). Informants are not all equal: predictors and correlates of clinician judgments about caregiver and youth credibility. *Journal of Child and Adolescent Psychopharmacology, 21*, 407-415. doi: 10.1089/cap.2011.0032

Youngstrom, E. A., Youngstrom, J. K., & Starr, M. (2005). Bipolar diagnoses in community mental health: Achenbach CBCL profiles and patterns of comorbidity. *Biological Psychiatry, 58,* 569-575.

Zhou, X.-H., Obuchowski, N. A., & McClish, D. K. (2002). *Statistical methods in diagnostic medicine*. New York, NY: Wiley.